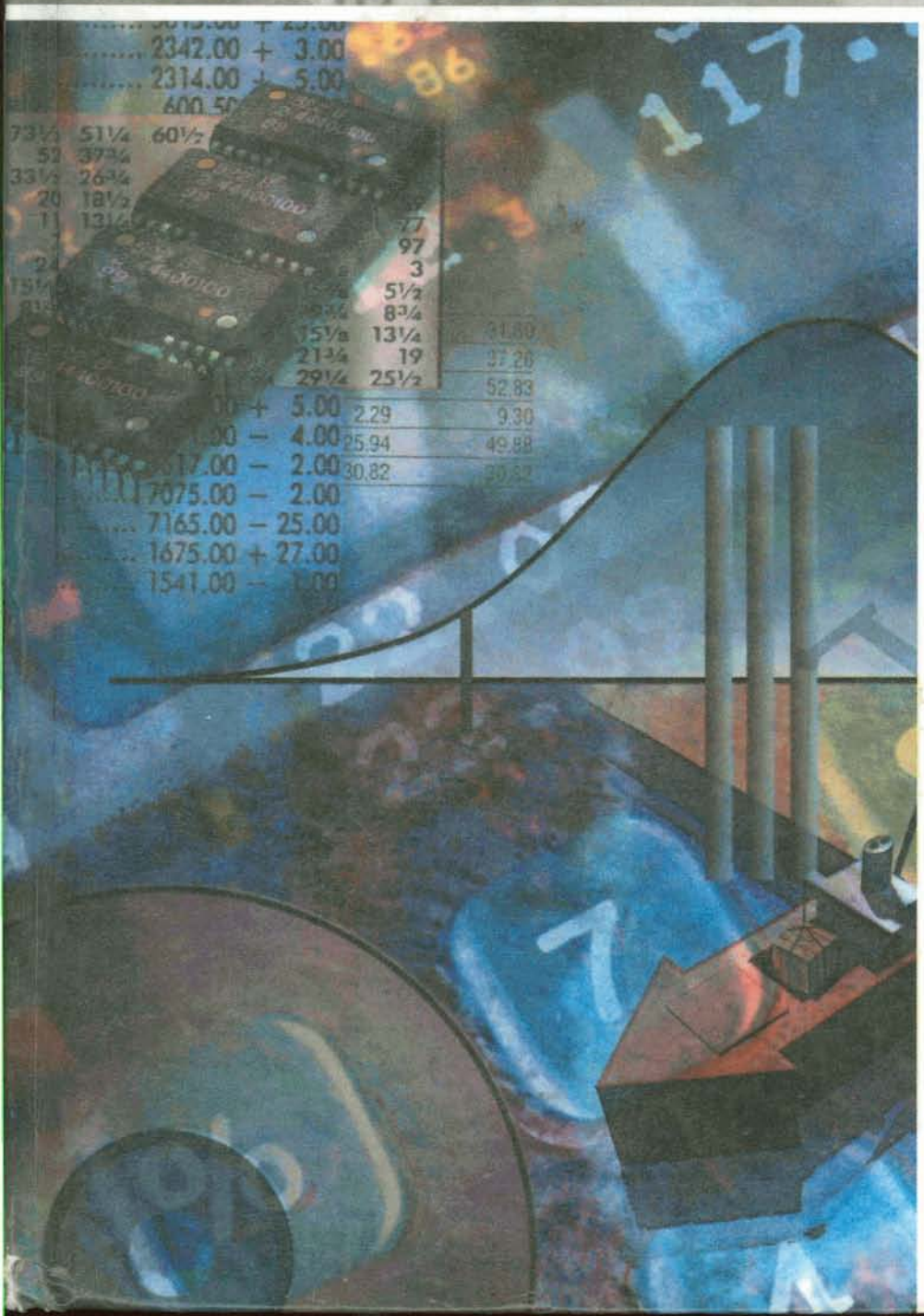
ПРАКТИЧЕСКАЯ БИЗНЕС-СТАТИСТИКА



ЧЕТВЕРТОЕ ИЗДАНИЕ

ЭНДРЮ Ф. СИГЕЛ

Практическая бизнес-статистика

Четвертое издание

Practical Business Statistics

FOURTH EDITION

ANDREW F. SIEGEL
Department of Management Science
Department of Finance
Department of Statistics
Department of Molecular Biotechnology
University of Washington

Irwin
McGraw-Hill

Boston Burr Ridge, IL Dubuque, IA Madison, WI New York San Francisco St. Louis
Bangkok Bogota Caracas Lisbon London Madrid
Mexico City Milan New Delhi Seoul Singapore Sydney Taipei Toronto

Практическая бизнес-статистика

Четвертое издание

ЭНДРЮ Ф. СИГЕЛ
Университет штата Вашингтон
Факультет научного менеджмента
Факультет финансов
Факультет статистики
Факультет молекулярной биотехнологии

ВИЛЬЯМС

Москва • Санкт-Петербург • Киев
2002

ББК 88.5$_я$75
С34
УДК 681.3.07

Издательский дом "Вильямс"

Зав. редакцией *С.Н. Тригуб*

Перевод с английского *А.И. Мороза, О.Л. Пелявского, А.В. Редины* и *Е.Л. Усенко*

Под редакцией канд. экон. наук. *А.П. Горбачика*

По общим вопросам обращайтесь в Издательский дом "Вильямс" по адресу:
info@williamspublishing.com, http://www.williamspublishing.com

Сигел, Эндрю.

С34 Практическая бизнес-статистика. : Пер. с англ. — М. : Издательский дом "Вильямс", 2002. — 1056 с. : ил. — Парал. тит. англ.

ISBN 5-8459-0306-8 (рус.)

Эта книга представляет собой прекрасно организованный вводный курс статистических методов анализа данных. Дидактически грамотно представленный теоретический материал не перегружен математическими подробностями и дополняется большим количеством тщательно отобранных примеров. Здесь есть анализ финансового состояния предприятий и конъюнктуры фондового рынка, прогнозирование уровня продаж и результатов избирательных кампаний, анализ качества продукции и эффективности рекламы, изучение аудитории средств массовой информации и много других непростых и практически важных задач. Книга может быть полезна преподавателям, студентам, научным сотрудникам, аналитикам консалтинговых фирм и рекламных агентств, всем тем, кто занимается (или еще только собирается заняться) прикладным статистическим анализом эмпирических экономических и социальных данных.

ББК 88.5$_я$75

Все названия программных продуктов являются зарегистрированными торговыми марками соответствующих фирм.

Никакая часть настоящего издания ни в каких целях не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, будь то электронные или механические, включая фотокопирование и запись на магнитный носитель, если на это нет письменного разрешения издательства *Irvin, McGraw-Hill*.

Authorized translation from the English language edition published by Irvin, McGraw-Hill Companies, Copyright © 2000 Andrew F. Siegel

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Russian language edition published by Williams Publishing House according to the Agreement with R&I Enterprises International, Copyright © 2002

ISBN 5-8459-0306-8 (рус.)
ISBN 0-07-117788-4 (англ.)

© Издательский дом "Вильямс", 2002
© Andrew F. Siegel, 2000

Оглавление

Предисловие 17

ЧАСТЬ I. ВВЕДЕНИЕ И ОПИСАТЕЛЬНАЯ СТАТИСТИКА 27

Глава 1. Введение: роль статистики в бизнесе 28

Глава 2. Структуры данных: классификация различных типов наборов данных 42

Глава 3. Гистограммы: взгляд на распределение данных 70

Глава 4. Обобщающие показатели: интерпретация типических значений и перцентилей 117

Глава 5. Изменчивость: изучение разнообразия 169

ЧАСТЬ II. ВЕРОЯТНОСТЬ 219

Глава 6. Вероятность: разбираемся в случайных ситуациях 220

Глава 7. Случайные величины: работа с неопределенными значениями 278

ЧАСТЬ III. СТАТИСТИЧЕСКИЙ ВЫВОД 339

Глава 8. Построение случайной выборки: предварительное планирование для сбора данных 340

Глава 9. Доверительные интервалы: допущение о неточности оценок 395

Глава 10. Проверка статистических гипотез: выбор между реальностью и совпадением 445

ЧАСТЬ IV. РЕГРЕССИЯ И ВРЕМЕННЫЕ РЯДЫ 517

Глава 11. Корреляция и регрессия: измерение и прогнозиро-вание взаимосвязей 518

Глава 12. Множественная регрессия: прогнозирование одного фактора на основе нескольких других 611

Глава 13. Составление отчетов: представление результатов множественной регрессии 720

Глава 14. Временные ряды: анализ изменений во времени 743

ЧАСТЬ V. МЕТОДЫ И ПРИМЕНЕНИЯ 807

Глава 15. Дисперсионный анализ: проверка различий для нескольких выборок и многое другое 808

Глава 16. Непараметрические методы: проверка гипотез для порядковых данных или данных, не подчиня-ющихся нормальному распределению 847

Глава 17. Анализ "хи-квадрат": поиск закономерностей для качественных данных 878
Глава 18. Контроль качества: выявление вариации и управление ею 908
Приложение А. База данных служащих 945
Приложение Б. Самопроверка: решение некоторых задач, а также упражнений, использующих базу данных 948
Приложение В. Статистические таблицы 976
Приложение Г. Краткое руководство по применению StatPad 1017
Словарь терминов 1025
Предметный указатель 1047

Содержание

Предисловие 17
Большая работа: пересмотр материала и проверка в студенческой аудитории 17
Стиль книги 18
Примеры 18
Статистические графики 18
Ситуации для анализа 19
Что нового в четвертом издании 19
Структура книги 19
Руководство к Excel 22
Не обязательное для изучения приложение StatPad 22
Руководство для преподавателя 22
Тесты 22
Благодарности 23
Обращение к студентам 24
Об авторе 25
ЧАСТЬ I. ВВЕДЕНИЕ И ОПИСАТЕЛЬНАЯ СТАТИСТИКА 27
Глава 1. Введение: роль статистики в бизнесе 28
1.1. Почему именно статистика? 28
Почему необходимо изучать статистику? 29
Сложна ли статистика? 29
Влияет ли знание статистики на гибкость принятия решений? 29
1.2. Что такое статистика? 29
Статистика рассматривает общую картину 30
Статистика не игнорирует отдельные объекты 30
Посмотрим на данные 31
Статистика в менеджменте 31
1.3. Четыре основных этапа статистического анализа 31
Оценка неизвестной величины 33
Проверка гипотез 34
1.4. Что такое вероятность 36
1.5. Общий совет 37
1.6. Дополнительный материал 37
Резюме 37
Основные термины 38
Контрольные вопросы 39
Задачи 39
Проект 41
Глава 2. Структуры данных: классификация различных типов наборов данных 42
2.1. Сколько переменных? 42
Одномерные данные 43

Двумерные данные 44
Многомерные данные 45
2.2. Количественные данные: числа 47
Дискретные количественные данные 47
Непрерывные количественные данные 47
Остерегайтесь чисел, не имеющих содержательную интерпретацию 48
2.3. Качественные данные: категории 49
Порядковые качественные данные 49
Номинальные качественные данные 50
2.4. Временные ряды и данные об одном временном срезе 50
2.5. Источники данных, включая Internet 52
2.6. Дополнительный материал 59
Резюме 59
Основные термины 61
Контрольные вопросы 61
Задачи 62
Упражнения с использованием базы данных 69
Проекты 69
Глава 3. Гистограммы: взгляд на распределение данных 70
3.1. Последовательность данных 71
Числовая ось 72
3.2. Использование гистограмм для отображения частот 73
Гистограммы и столбиковые диаграммы 77
3.3. Нормальное распределение 79
3.4. Несимметричные распределения и преобразование данных 81
Проблема с асимметрией 85
Выход с помощью преобразования 85
Интерпретация и вычисление логарифма 87
3.5. Бимодальные распределения 88
3.6. Выбросы (сильно отклоняющиеся значения) 91
Работа с выбросами (сильно отклоняющимися значениями) 92
3.7. Гистограммы, построенные вручную: метод "ствол и листья" 97
3.8. Дополнительный материал 99
Резюме 99
Основные термины 101
Контрольные вопросы 101
Задачи 102
Упражнения с использованием базы данных 114
Проекты 115
Ситуация для анализа 115
Необходимость контроля производственных потерь 115
Вопросы для обсуждения 116
Глава 4. Обобщающие показатели: интерпретация типических значений и перцентилей 117
4.1. Чему равно наиболее типическое значение? 118
Среднее: типическое значение для количественных данных 118
Взвешенное среднее: учет важности 121
Медиана: типическое значение для количественных и порядковых данных 126
Мода: типическое значение даже для номинальных данных 133
Какие показатели нужно использовать 135
4.2. Что такое перцентиль 137
Экстремумы, квартили и блочные диаграммы 137
Функция кумулятивного распределения показывает перцентили 142
4.3. Дополнительный материал 149

Резюме 149
Основные термины 151
Контрольные вопросы 152
Задачи 153
Упражнения с использованием базы данных 164
Проекты 164
Ситуация для анализа 165
Управленческие прогнозы о производстве и маркетинге, или "Случай подозрительного потребителя" 165
Вопросы для обсуждения 168
Глава 5. Изменчивость: изучение разнообразия 169
5.1. Стандартное отклонение: традиционный выбор 171
Определение и формула для стандартного отклонения и дисперсии 172
Использование калькулятора или компьютера 174
Интерпретация стандартного отклонения 175
Интерпретация стандартного отклонения для нормального распределения 177
Стандартное отклонение выборки и генеральной совокупности 187
5.2. Размах: быстрая и поверхностная оценка 189
5.3. Коэффициент вариации: мера *относительной* изменчивости 192
5.4. Результаты прибавления константы или изменения шкалы 194
5.5. Дополнительный материал 196
Резюме 196
Основные термины 198
Контрольные вопросы 199
Задачи 200
Упражнения с использованием базы данных 215
Проекты 216
Ситуация для анализа 216
Следует ли продолжать работу с этим поставщиком? 216
ЧАСТЬ II. ВЕРОЯТНОСТЬ 219
Глава 6. Вероятность: разбираемся в случайных ситуациях 220
6.1. Пример: за какой из дверей спрятан приз? 222
6.2. Как исследовать неопределенность 223
Случайный эксперимент: *точное* определение случайной ситуации 223
Выборочное пространство: перечень возможных событий 224
Результат: что происходит в действительности 226
События: они либо происходят, либо нет 227
6.3. Насколько вероятно событие? 229
Каждое событие имеет свою вероятность 229
Откуда берутся значения вероятности 230
Относительная частота и закон больших чисел 230
Теоретическое значение вероятности 233
Правило равной вероятности 233
Субъективная оценка вероятности 234
Анализ методом Байеса и частотный анализ 236
6.4. Как совместить информацию о нескольких событиях 237
Диаграммы Венна позволяют увидеть все возможности 237
Не событие 238
Правило дополнения (правило *не*) 238
Одно событие *и* другое 239
Если два события не могут наблюдаться вместе 240
Правило пересечения (*и*) для несовместимых событий 240
Одно событие *или* другое 240
Правило объединения (*или*) для несовместимых событий 241
Нахождение *или* из *и* и наоборот 242

Одно событие *при условии* другого: учет имеющейся информации 243
Правило вычисления условной вероятности при наличии дополнительной информации 244
Условные вероятности *для несовместимых* событий 245
Независимые события 246
Правило пересечения (*и*) для независимых событий 248
Связь между независимыми и несовместимыми событиями 249
6.5. Как решать вероятностные задачи 249
Дерево вероятностей 249
Правила построения дерева вероятностей 250
Таблица совместных вероятностей 257
6.6. Дополнительный материал 259
Резюме 259
Основные термины 262
Контрольные вопросы 263
Задачи 264
Упражнения с использованием базы данных 274
Проект 275
Ситуация для анализа 276
Детективная история: кто же все-таки ответствен за увеличение количества дефектов в последнее время? 276
Глава 7. Случайные величины: работа с неопределенными значениями 278
7.1. Дискретные случайные величины 280
Вычисление среднего и стандартного отклонения 281
7.2. Биномиальное распределение 285
Определение биномиального распределения и биномиального соотношения 285
Вычисление среднего и стандартного отклонения: короткий путь 288
Вычисление вероятностей 290
7.3. Нормальное распределение 295
Представление вероятности как площади под кривой 297
Стандартное нормальное распределение Z и соответствующие вероятности 298
Решение задач на вычисление вероятности при нормальном распределении 302
Четыре способа вычисления вероятности 308
Внимание! Не все распределено нормально! 308
7.4. Аппроксимация биномиального распределения нормальным 309
7.5. Распределение Пуассона и экспоненциальное распределение 315
Распределение Пуассона 315
Экспоненциальное распределение 318
7.6. Дополнительный материал 320
Резюме 320
Основные термины 323
Контрольные вопросы 324
Задачи 325
Упражнения с использованием базы данных 335
Проекты 335
Ситуация для анализа 336
Стоимость опциона на аренду нефтяного месторождения 336
ЧАСТЬ III. СТАТИСТИЧЕСКИЙ ВЫВОД 339
Глава 8. Построение случайной выборки: предварительное планирование для сбора данных 340
8.1. Генеральные совокупности и выборки 341
Что такое репрезентативная выборка 343

Параметры выборки и параметры генеральной совокупности 345
8.2. Случайная выборка 346
Извлечение случайной выборки 347
Извлечение выборки методом перемешивания генеральной совокупности 350
8.3. Выборочное распределение и центральная предельная теорема 352
8.4. Стандартная ошибка как *оценка* стандартного отклонения 359
Насколько отличается среднее выборки от среднего генеральной совокупности 360
Поправка для малой генеральной совокупности 363
Стандартная ошибка биномиальной доли 366
8.5. Другие методы построения выборки 367
Стратифицированная случайная выборка 368
Систематическая выборка 372
8.6. Дополнительный материал 374
Резюме 374
Основные термины 378
Контрольные вопросы 378
Задачи 380
Упражнения с использованием базы данных 390
Проекты 392
Ситуация для анализа 393
Можно ли извлечь пользу из этого исследования? 393
Глава 9. Доверительные интервалы: допущение о неточности оценок 395
9.1. Доверительный интервал для среднего значения и для доли признака в генеральной совокупности 398
t-таблица и t-распределение 401
Часто используемый 95% доверительный интервал 404
Другие доверительные уровни 409
9.2. Предположения, необходимые для корректного использования 412
Случайная выборка 413
Нормальное распределение 416
9.3. Интерпретация доверительного интервала 417
Какое событие имеет вероятность 95%? 417
Ваши жизненные достижения 418
9.4. Односторонние доверительные интервалы 419
Внимание! Не всегда можно использовать односторонний доверительный интервал 419
Вычисление одностороннего доверительного интервала 420
9.5. Интервалы предсказания 423
9.6. Дополнительный материал 426
Резюме 426
Основные термины 428
Контрольные вопросы 428
Задачи 430
Упражнения с использованием базы данных 440
Проекты 442
Ситуация для анализа 443
Многообещающие результаты опроса относительно заказов фирменных товаров по каталогу 443
Глава 10. Проверка статистических гипотез: выбор между реальностью и совпадением 445
10.1. Не все гипотезы одинаковы! 446
Нулевая гипотеза 446
Исследовательская гипотеза 447

О чем свидетельствует результат 447
Примеры гипотез 448
10.2. Проверка гипотезы о равенстве среднего генеральной совокупности некоторому заданному значению 450
Использование доверительных интервалов: простой способ 451
t-статистика: способ другой, результат тот же 459
10.3. Интерпретация проверки гипотезы 462
Ошибки I и II рода 462
Условия применимости 464
Гипотезы не могут быть вероятно истинными или вероятно ложными 464
Статистическая значимость и уровни проверки 465
Доверительная вероятность (p-значение) 466
10.4. Односторонняя проверка 469
Как выполнять проверку 471
10.5.4. Проверка того, принадлежит ли новое наблюдение той же генеральной совокупности 478
10.6. Сравнение двух выборок 479
t-тест для зависимых выборок 480
t-тест для независимых выборок 482
10.7. Дополнительный материал 487
Резюме 487
Основные термины 492
Контрольные вопросы 493
Задачи 495
Упражнения с использованием базы данных 512
Проекты 513
Ситуация для анализа 513
Так много рекламы, так мало времени 513

ЧАСТЬ IV. РЕГРЕССИЯ И ВРЕМЕННЫЕ РЯДЫ 517

Глава 11. Корреляция и регрессия: измерение и прогнозиро-вание взаимосвязей 518

11.1. Исследование взаимосвязей с помощью диаграмм рассеяния и корреляций 519
Диаграмма рассеяния демонстрирует взаимосвязь 520
Корреляция характеризует силу взаимосвязи 521
Формула для вычисления коэффициента корреляции 522
Различные типы взаимосвязей 525
Отсутствие взаимосвязи 531
Нелинейная взаимосвязь 533
Неодинаковая вариация 536
Разделение совокупности на группы 540
Выбросы (резко отклоняющиеся значения) 542
Корреляция — это не причинная обусловленность 544
11.2. Регрессия: предсказание одного фактора на основании другого 546
Прямая линия характеризует линейную взаимосвязь 547
Прямые линии 549
Построение линии на основе данных 549
Насколько полезна построенная линия 555
Стандартная ошибка оценки: насколько велики ошибки предсказания 555
R^2: как много объяснено 558
Доверительные интервалы и проверка гипотез для регрессии 558
Предположение о линейности определяет генеральную совокупность 559
Стандартные ошибки для наклона и сдвига 560
Доверительные интервалы для коэффициентов регрессии 561
Проверка того, является связь реальной или случайной 562
Другие методы проверки значимости взаимосвязи 563

Результаты компьютерных вычислений для данных о производственных затратах 564
Проверки других гипотез о коэффициенте регрессии 566
Новое наблюдение: неопределенность и доверительный интервал 568
Среднее значение Y: неопределенность и доверительный интервал 570
Регрессия может вводить в заблуждение 572
Линейная модель может оказаться неверной 573
Трудно предсказать вмешательство исходя из наблюдаемого опыта 576
Сдвиг может быть лишен смысла 576
Представление Y на основании X и представление X на основании Y 577
Скрытый "третий фактор" может быть полезен 578
11.3. Дополнительный материал 578
Резюме 578
Основные термины 583
Контрольные вопросы 583
Задачи 586
Упражнения с использованием базы данных 605
Проекты 607
Ситуация для анализа 608
Еще один этап производства: нужен ли он? 608

Глава 12. Множественная регрессия: прогнозирование одного фактора на основе нескольких других 611

12.1. Интерпретация результатов множественной регрессии 613
Коэффициенты регрессии и уравнение регрессии 618
Интерпретация коэффициентов регрессии 620
Прогнозы и ошибки прогнозирования 621
Насколько хороши наши прогнозы 625
Типичная ошибка прогнозирования: стандартная ошибка оценки 625
Объясненный процент вариации: R^2 625
Статистический вывод в случае множественной регрессии 626
Предположения 626
Значима ли модель? F-тест или тест R^2 628
Таблицы критических значений для тестирования R^2 632
Какие переменные являются значимыми: t-тест для каждого коэффициента 641
Другие проверки, касающиеся коэффициента регрессии 644
Какие переменные оказывают большее влияние 645
Сравнение стандартизованных коэффициентов регрессии 645
Сравнение коэффициентов корреляции 647
12.2. Сложности и проблемы, связанные с множественной регрессией 649
Мультиколлинеарность: не слишком ли схожи между собой объясняющие переменные? 650
Выбор переменной: может быть, мы пользуемся "не теми" переменными? 656
Классификация перечня X-переменных по приоритетам 657
Автоматизация процесса выбора переменных 658
Неправильный выбор модели: возможно, уравнение регрессии имеет неправильную форму? 660
Анализ данных с целью выявления нелинейности или неравной изменчивости 661
Использование диагностической диаграммы для выяснения наличия проблем 662
Использование процентных изменений для моделирования экономических временных рядов 669
12.3. Нелинейные взаимосвязи и неравная изменчивость 672
Преобразование взаимосвязи в линейную форму: интерпретация результатов 673
Подгонка кривой с помощью полиномиальной регрессии 679
Моделирование взаимодействия между двумя X-переменными 682

12.4. Индикаторные переменные: прогнозирование на основе категорий 684
Интерпретация и проверка значимости коэффициентов регрессии для индикаторных переменных 686
Раздельные регрессии 691
12.5. Дополнительный материал 692
Резюме 692
Основные термины 695
Контрольные вопросы 696
Задачи 697
Упражнения с использованием базы данных 714
Проект 716
Ситуация для анализа 717
Контроль качества продукции 717
Глава 13. Составление отчетов: представление результатов множественной регрессии 720
13.1. Как организовать свой отчет 722
Абзац, содержащий реферат 723
Вводная часть 724
Раздел "Анализ и методы" 724
Раздел "Выводы и резюме" 726
Включение ссылок 726
Раздел приложений 728
13.2. Рекомендации и советы 728
Помните о своей аудитории 728
О чем писать в первую очередь, во вторую, в последнюю? 729
Другие источники 729
13.3. Пример: формула оперативного ценообра-зования для ответа на запросы потребителей 730
13.4. Дополнительный материал 736
Резюме 736
Основные термины 738
Контрольные вопросы 738
Задачи 739
Упражнение с использованием базы данных 741
Проект 741
Глава 14. Временные ряды: анализ изменений во времени 743
14.1. Обзор анализа временных рядов 744
14.2. Анализ трендов и сезонности 755
Тренд и циклический компонент: скользящее среднее 757
Сезонный индекс: среднее значение отношения к скользящему среднему отражает сезонное поведение 759
Поправка на сезон: деление ряда на сезонный индекс 763
Долгосрочный тренд и прогноз с поправкой на сезонные колебания: линия регрессии 766
Прогноз: тренд с учетом сезонности 770
14.3. Моделирование циклического поведения с помощью ARIMA-процессов Бокса–Дженкинса 771
Процесс случайного шума не обладает памятью: отправная точка 774
Процесс авторегрессии (AR) обладает памятью о своем прошлом 775
Процесс скользящего среднего (MA) имеет ограниченную память 776
Процесс авторегрессии и скользящего среднего (ARMA) сочетает в себе AR и MA 778
Чистый интегрированный (I) процесс помнит, где он находился, и затем движется случайно 782
Процесс авторегрессионного интегрированного скользящего среднего (ARIMA) помнит свои изменения 785

14.4. Дополнительный материал 787
Резюме 787
Основные термины 791
Контрольные вопросы 792
Задачи 794
Проекты 804
ЧАСТЬ V. МЕТОДЫ И ПРИМЕНЕНИЯ 807
Глава 15. Дисперсионный анализ: проверка различий для нескольких выборок и многое другое 808
15.1. Использование блочных диаграмм для одновременного представления нескольких выборок 810
15.2. F-тест определяет, значимо ли различаются средние 812
Данные и источники вариации 812
Допущения 813
Гипотезы 814
F-статистика 815
F-таблица 817
Результат F-теста 822
Результат вычислений с помощью компьютера: однофакторная ANOVA-таблица 823
15.3. Тест наименьшего значимого различия: какие пары различаются? 824
15.4. Более сложные планы дисперсионного анализа 827
Разнообразие — вот, что придает вкус жизни 828
Двухфакторный дисперсионный анализ 828
Три фактора и более 830
Ковариационный анализ, ANCOVA 830
Многомерный дисперсионный анализ (MANOVA) 830
Как читать таблицу ANOVA 830
15.5. Дополнительный материал 834
Резюме 834
Основные термины 836
Контрольные вопросы 837
Задачи 837
Упражнения с использованием базы данных 845
Проекты 846
Глава 16. Непараметрические методы: проверка гипотез для порядковых данных или данных, не подчиня-ющихся нормальному распределению 847
16.1. Проверка гипотезы о равенстве медианы некоторому заданному значению 849
Критерий знаков 850
Гипотезы 850
Допущение 851
16.2. Тестирование различий в двух связанных выборках 855
Использование критерия знаков для разностей 855
Гипотезы 856
Условие 857
16.3. Проверка значимости различия двух независимых выборок 858
Процедура, основанная на ранжировании *всех* данных 858
Гипотезы 859
Допущения 860
16.4. Дополнительный материал 864
Резюме 864
Основные термины 867

Контрольные вопросы 868
Задачи 869
Упражнения с использованием базы данных 876
Проекты 876

Глава 17. Анализ "хи-квадрат": поиск закономерностей для качественных данных 878

17.1. Обобщение качественных данных с помощью частот и процентов 879
17.2. Проверка того, что значения процентов в генеральной совокупности равны некоторым заданным значениям 881
Критерий "хи-квадрат" в отношении равенства процентов 881
17.3. Проверка взаимосвязи между двумя качественными переменными 888
Понятие независимости переменных 888
Критерий "хи-квадрат" независимости 889
17.4. Дополнительный материал 895
Резюме 895
Основные термины 897
Контрольные вопросы 897
Задачи 898
Упражнения с использованием базы данных 905
Проекты 906

Глава 18. Контроль качества: выявление вариации и управление ею 908

18.1. Процессы и причины вариации 911
Диаграмма Парето показывает, на что обратить внимание 913
18.2. Что такое карты контроля и как их читать 915
Контрольные границы показывают выход из-под контроля одного наблюдения 916
Как выявить проблему даже в пределах контрольных границ 917
18.3. Отображение количественных измерений в $\bar{X}$- и R-картах 919
18.4. Построение карт контроля для процента брака 926
18.5. Дополнительный материал 930
Резюме 930
Основные термины 932
Контрольные вопросы 932
Задачи 933
Проекты 944

Приложение А. База данных служащих 945

Приложение Б. Самопроверка: решение некоторых задач, а также упражнений, использующих базу данных 948

Приложение В. Статистические таблицы 976

Приложение Г. Краткое руководство по применению StatPad 1017

Словарь терминов 1025

Предметный указатель 1047

Предисловие

Традиционный курс бизнес-статистики изменился и в основном к лучшему: больше внимания уделяют интерпретации данных, понятиям и идеям, лучше демонстрируют взаимосвязь статистики с различными видами экономической деятельности, все делается для более глубокого понимания фундаментальных статистических принципов. Книга *Практическая бизнес-статистика* (*Practical Business Statistics*) является лидером этих перемен. В предисловии к первым трем изданиям говорилось следующее.

"Традиционный курс бизнес-статистики меняется. В настоящее время в связи с доступностью и широким использованием компьютеров для обработки числовой информации нет необходимости в подробном освещении многих старомодных тем. Это дало огромную возможность для подачи в учебное время нового материала, который необходим бизнес-менеджеру, – нового материала, который подводит фундамент под понятия и приложения статистики применительно к бизнесу и экономике. Например, менеджерам не обязательно знать, как вывести формулу для вычисления коэффициентов регрессии в методе наименьших квадратов, но они должны уметь интерпретировать коэффициенты регрессии, чтобы, опираясь на эти важные показатели, принимать верные решения в сложных ситуациях.

Практическая бизнес-статистика была написана с учетом этих изменений. Студенты экономических специальностей могут выучить статистику на "отлично", если изучение статистики сопровождать реальными практическими примерами и простыми и доступными объяснениями, из которых видно, почему стоит осваивать статистический взгляд на мир. Студенты-экономисты отличаются от студентов других специальностей и заслуживают книги, созданной специально в соответствии с их потребностями и интересами."

Большая работа: пересмотр материала и проверка в студенческой аудитории

В начале работы над этой книгой был сборник материалов для самостоятельного чтения, который я раздал своим студентам в дополнение к рекомендуемому учебнику. Все доступные книги представляли статистику излишне сложно, а я хотел представить прямые и легкие способы понимания предмета. Я также хотел придать предмету аромат современной коммерческой деятельности. Вся полезная информация, которую в качестве обратной связи я получил от студентов за все эти годы, была переработана и учтена в книге.

Даже перед публикацией первого издания *Практическая бизнес-статистика* прошла несколько стадий рецензирования и проверки в студенческой аудитории. Сейчас, когда в университетах и колледжах всей страны да и всего мира используют три предыдущих издания, подготовка четвертого издания дала мне возможность улучшить книгу с учетом всех полученных мною дополнительных рецензий и полезных положительных комментариев.

Стиль книги

Мне нравилось писать эту книгу. Везде, где это было возможно, *я показывал* на примерах "область применения" и объяснял, как мы, статистики, *действительно* размышляем о предмете, что это означает и насколько это полезно. Такой подход помог мне вдохнуть жизнь в предмет, который, к сожалению, имеет непривлекательный образ в глазах многих людей. Конечно, в книге даны также и традиционные объяснения, что позволяет иметь два взгляда на вещи: вот, что мы говорим, и вот, что это означает.

Я был взволнован, услышав, когда некоторые из моих студентов, которые просто боялись математики, сказали, что книгу действительно *приятно читать*! И это *после* того, как они получили свои оценки на экзамене!

Примеры

Примеры приближают статистику к жизни, показывают полезность и важность каждой темы. Книга *Практическая бизнес-статистика* содержит много реальных примеров, отобранных из широкого круга экономических источников. Фондовый рынок рассматривается в главе 5 для демонстрации рыночной изменчивости и риска, измеренных с помощью стандартного отклонения. Контроль качества рассматривается в разных главах для иллюстрации отдельных тем, а также в главе 18, которая посвящена именно этой теме. Опросы общественного мнения и предвыборные опросы также рассматриваются в разных главах (и особенно в главе 9), поскольку они представляют в чистом виде статистический вывод для реальной, хорошо знакомой всем нам, жизненной ситуации. Поиск данных в Internet рассмотрен в главе 2. На примере цен на рекламу в главе 12 показано, как множественная регрессия может выявить взаимосвязи в сложном наборе данных. Для демонстрации прогнозирования на основе временных рядов в главе 14 рассмотрены данные об уровне безработицы. Использование наглядных примеров значительно облегчает обучение студентов.

Статистические графики

Чтобы наглядно показать, что происходит с данными, *Практическая бизнес-статистика* включает более 200 рисунков, иллюстрирующих основные характеристики и зависимости. Эти графики точны, поскольку выполнены с помощью компьютера. Например, имеющая форму колокола кривая нормального распределения здесь представлена точно, в отличие от аналогичных кривых, изображенных в некоторых книгах, где эта форма непроизвольно улучшена художником. Но точность нельзя заменить ничем! Это помогает студентам лучше понять и запомнить статистические понятия.

Ситуации для анализа

В конце глав 3–12 помещены ситуационные задачи, которые демонстрируют полезность статистического мышления как неотъемлемой части более широкой коммерческой деятельности. Эти задачи часто не имеют единственно верного решения, но дают возможность для размышлений и открытых обсуждений.

Что нового в четвертом издании



Здесь подытожены основные изменения. Многие примеры и задачи взяты из предыдущих изданий, но теперь они основаны на более свежих данных. Книга проиллюстрирована копиями экрана Excel, которые показывают, как выполнять многие операции непосредственно с помощью электронной таблицы. В главе 2 рассматриваются новые источники данных, в частности Internet. Глобальная сеть Internet также часто предлагается в качестве средства поиска данных для проектов, представленных в конце каждой главы. Связанные с Internet задачи или источники данных помечены в книге специальной пиктограммой. Внося изменения, я старался сохранить те части книги из первых трех изданий, которые хорошо зарекомендовали себя среди студентов и преподавателей.

Структура книги

Студент всегда должен знать, *почему* предлагаемый материал является важным. Поэтому каждая часть книги начинается с краткого обсуждения рассматриваемого предмета и соответствующих глав. Каждая глава начинается с краткого обзора основной темы, который демонстрирует важность этой темы для бизнеса, прежде чем будут изложены детали и примеры.

Ключевые слова, основные термины и понятия выделены жирным шрифтом. Они собраны в списке основных понятий в конце каждой главы, а также включены в словарь терминов. Это облегчает их изучение и дает возможность сконцентрировать внимание на основных идеях. Предметный указатель поможет легко и быстро найти как главные темы, так и отдельные детали. Например, попытайтесь найти термины "корреляция", "временной ряд" или "доверительный интервал".

Каждая глава заканчивается разделом "Краткое содержание и задачи", который начинается с резюме содержания основного материала. Затем идет перечень основных терминов. Контрольные вопросы дают обзор основных тем, указывая также на их важность. Предложенные *задачи* дают студенту возможность применить статистику в новых ситуациях. "Упражнения с использованием базы данных" предлагают также ряд практических задач с использованием данных о служащих из приложения А. "Проекты" призваны приблизить статистику к потребностям и интересам студентов. Студенты могут самостоятельно ставить задачу и, руководствуясь своим опытом или интересами, подбирать данные из различных источников, включая Internet, различные публикации или отчеты компаний. И наконец, "Ситуации для анализа" (по одной в каждой из глав 3–12) дают возможность для размышлений и обсуждений, часто без единственно верного ответа.

В дополнение к достаточно традиционным основам статистики и применению ее в бизнесе представлено также несколько относительно новых тем. Учитывая большое значение для бизнеса обмена информацией, включена глава 13, посвященная сбору статистических материалов и представлению их в виде отчета. Глава 14 включает наглядное обсуждение метода Бокса–Дженкинса прогнозирования временных рядов на основе моделей ARIMA. Глава 18 демонстрирует, как статистические методы помогают в повышении качества продукции; обсуждение методов контроля качества можно встретить также и в других главах книги.

Книга *Практическая бизнес-статистика* состоит из пяти частей и четырех приложений.

Часть I, "Введение и описательная статистика" (главы 1–5). Глава 1 пробуждает интерес у студента, показывая, как использование статистики обеспечивает конкурентное преимущество в бизнесе. В главе 2 представлен обзор различных типов данных (количественные, качественные, порядковые, номинальные, двумерные, временные ряды и т.п.), показана разница между первичными и вторичными данными, а также рассматривается использование Internet. В главе 3 показано, как гистограммы позволяют обнаружить такие особенности данных, которые сложно определить просто из колонок чисел. Глава 4 охватывает такие базовые статистические показатели, как среднее, медиана, мода и перцентили, которые отображаются на блочных диаграммах и графиках функций распределения. В главе 5 в терминах стандартного отклонения, размаха и коэффициента вариации обсуждается понятие изменчивости, которое в бизнесе часто называют *риском*.

Часть II, "Вероятность" (главы 6, 7). В главе 6 рассмотрены вероятности событий и их комбинаций, а также использование дерева вероятностей как для визуализации (наглядного представления) ситуации, так и для вычисления вероятностей. Условные вероятности интерпретируют как подход к наилучшему использованию имеющейся информации. В главе 7 рассматриваются случайные переменные (результаты, представленные в виде чисел), которые часто представляют числовые значения показателей, важных для вашего бизнеса, но непосредственно недоступных. Подробно рассмотрены дискретное распределение в общем виде, нормальное распределение, биномиальное распределение, а также (кратко) экспоненциальное распределение и распределение Пуассона.

Часть III, "Статистический вывод" (главы 8–10). Эти главы сводят вместе описательные показатели из части I и формальные вероятностные оценки из части II. В главе 8 рассматриваются процесс взятия случайной выборки, что является основой точного вероятностного утверждения статистического вывода, а также центральная предельная теорема и чрезвычайно важное понятие стандартной ошибки статистического показателя. В главе 9 демонстрируется, как, исходя из статистических данных, доверительные интервалы позволяют формулировать точные вероятностные утверждения о значении неизвестной величины. Рассмотрены одно- и двухсторонний доверительные интервалы, а также интервал предсказания для нового наблюдения. В главе 10 проверка статистических гипотез рассмотрена в значительной мере с точки зрения поиска отличий между

действительно существующими закономерностями в данных и случайными совпадениями. Применение представленного в главе 9 наглядного процесса построения доверительных интервалов делает проверку статистических гипотез интуитивно понятной и логически относительно несложной, сохраняя при этом строгую статистическую корректность.

Часть IV, "Регрессия и временные ряды" (главы 11–14). В этих главах понятия и методы из предыдущих частей книги применяют к более сложным и реальным ситуациям. Глава 12 демонстрирует изучение отношений между переменными и прогнозирование значений переменных на основе корреляций и регрессий для двумерных данных. Развитие такого подхода приводит к представленному в главе 12, вероятно, одному из наиболее важных методов статистики – методу множественной регрессии. В рассмотрении множественной регрессии особое внимание уделяется интерпретации результатов, диагностике и идее "контролирования" или "внесения поправок на влияние" одних факторов при изучении влияния других. Глава 13 содержит руководство к написанию отчетов, чтобы помочь студентам представлять результаты множественного регрессионного анализа для тех, кто занимается бизнесом. В главе 14 содержатся различные, в том числе и новые, методы, необходимые для анализа временных рядов. Метод прогнозирования трендов с поправкой на сезонные колебания используют, чтобы получить наглядное представление основных свойств временного ряда. В этой же главе изложены сложные современные методы Бокса–Дженкинса, позволяющие осуществлять прогнозирование в более трудных ситуациях.

Часть V, "Методы и применения" (главы 15–18), представляет собой совокупность необязательных (факультативных) специальных тем, которые значительно дополняют основной материал, представленный в предыдущих главах. В главе 15 показано, как с помощью дисперсионного анализа можно выполнить проверку гипотез для более сложных ситуаций. Глава 16 охватывает непараметрические методы, которые можно использовать в ситуациях, когда не выполняются основные предварительные условия для проверки гипотез, т.е. отсутствует нормальное распределение или речь идет о порядковых данных. В главе 17 показано, как можно использовать "хи-квадрат" анализ для проверки связи между категориями номинальных данных. Наконец, в главе 18 показано, как контроль качества в значительной мере связан с такими статистическими методами, как диаграммы Парето и карты контроля.

Представленная в приложении А база данных содержит информацию о заработной плате, стаже работы, возрасте, поле и об уровне квалификации некоторого количества административных служащих. Этот набор данных используется в разделах, посвященных упражнениям с базой данных, которые содержатся в конце большинства глав. В приложении Б представлены подробные решения некоторых (отмеченных в тексте пиктограммой) задач и упражнений с базой данных. В приложении В приведены все используемые в книге статистические таблицы. Приложение Г содержит краткое справочное руководство к StatPad-приложению, реализованному в среде Excel.

Руководство к Excel

В руководстве к Excel, подготовленном Эндрю Ф. Сигелом, приведены примеры статистического анализа данных из всех глав с помощью Excel. Это удобный способ научить студентов использовать для статистического анализа компьютер, если вы остановили свой выбор на Excel. Руководство находится на компакт-диске, который прилагается к этой книге.

Не обязательное для изучения приложение StatPad

StatPad является реализованным в среде Excel приложением, которое упрощает использование Excel для статистического анализа. Это приложение находится на компакт-диске, который прилагается к этой книге. Статистические методы организованы так, чтобы их можно было использовать легко и быстро, результаты выведены в рабочую таблицу вместе с соответствующими объяснениями. Кроме того, что StatPad упрощает выполнение вычислений, его использование гарантирует, что студенты будут иметь удобный доступ к статистическим методам и после завершения данного курса, и даже после окончания учебы! В приложении Г представлен краткий обзор возможностей StatPad.

StatPad был разработан Skyline Technologies, Inc. при участии Эндрю Ф. Сигела. Использование обычных диалоговых окон в стиле Windows и Excel, вывод результатов в электронную таблицу — все это делает статистические операции частью рабочей таблицы в такой же мере, как любые другие команды электронных таблиц. Эффективное использование компьютера позволяет быстрее и легче изучить основные понятия статистики!

Руководство для преподавателя

Руководство для преподавателя (ISBN 0-07-233611-0) разработано с целью помочь сократить время подготовки лекций. Каждая глава этого руководства снабжена коротким обзором изучаемой темы и рекомендациями о том, как заинтересовать студентов. Кроме того, в него включены подробные ответы на вопросы, решения задач и упражнений с базой данных, а также анализ и дискуссионный материал для ситуационных задач.

Тесты

Подготовленный Тедом Тзукахарой (Ted Tsukahara) набор тестов (ISBN 0-07-233612-9) содержит более 800 вопросов и задач. Вопросы упорядочены по степени трудности и соответствуют определенным главам книги. Есть также набор тестов для системы Diploma, которая представляет собой компьютерную программу для генерации тестов, позволяющую в среде Windows получить доступ и выбрать любой из тестов набора (ISBN 0-07-233615-3).

Благодарности

Я благодарен всем рецензентам и студентам, которые прочитали и прокомментировали черновик и предыдущие издания *Практической бизнес-статистики*. Я благодарен тем внимательным читателям, которые не побоялись прямо заявить о том, что необходимо сделать для улучшения книги с их точки зрения.

Я счастлив (и горжусь), что у меня была возможность постоянно общаться с моими родителями Милдред и Армандом Сигел (Mildred and Armand Siegel), которые дали мне ряд дельных советов относительно этой книги.

Передаю также слова благодарности Майклу Антонуччи (Michael Antonucci), моему местному торговому представителю, который подал идею этой книги, когда зашел ко мне в офис поговорить о компьютерах и посмотреть, чем я занимаюсь. Именно Майкл познакомил меня со многими хорошими людьми из издательства *Irwin*. До сих пор я встречаюсь со многими из них и все еще никак не могу понять, как *Irwin* удалось найти и вовлечь в работу столько прекрасных людей.

Я благодарен судьбе, что мне довелось сотрудничать с такими редакторами, как Скот Изенберг (Scott Isenberg), Гейл Короза (Gail Korosa) (редактор по развитию) и Денис Сантор-Митзит (Denis Santor-Midzit) (редактор проектов), которые вели этот проект. Я очень ценю помощь редакторов Ричарда Т. Херчера (Richard T. Hercher), Кэрол Роуз (Carol Rose), Эн Гранаки (Ann Granacki), Колин Тюшер (Collin Tusher) и Маргарет Хейвуд (Margaret Haywood), оказанную при подготовке первых изданий книги. Подготовка этой книги потребовала больших усилий, и я рад, что рядом со мной были самоотверженные люди, обладающие обширными знаниями и организаторскими способностями.

Творческая работа по дизайну текста выполнена Синтией Кремптон (Cynthia Crampton). Благодаря ей текст имеет не только эстетически привлекательный вид, но и упрощает поиск нужного материала. Мэри Христиансон (Mary Christianson) руководила художественной и дизайнерской работой, добившись согласованности разнообразных стилевых элементов, чтобы книга была не только функциональной, но и привлекательной.

Выражаю дополнительную благодарность за полезные комментарии Теду Тзукахара, который проверил точность всего текста книги и руководства преподавателю. Я благодарен Дэвиду Ауэру, который подготовил практические контрольные опросы на CD-ROM для студентов, а также Эрику Расселу, Дейтону Робинсону, Эрику Дж. Бину, Мишелю Р. Фанчеру, Сюзанне Степлетон, Саре С. Хемпфил, Ненси Дж. Силберг, А. Рональду Хауверу, Хирокуни Тамура, Джону Чу, Джун Морита, Брайан Макмуллену, Дэвиду Б. Фостеру, Пабло Ферреро, Рольфу Р. Андерсону, Гордону Клюгу, Е. Н. Функу, Робу Джуллетте, Девиду Хартнетт, Мики Лассу, Джудиан Морган, Кимберли В. Орчард, Ричарду Ричингсу, Марку Роеллигу, Скотту Х. Паттисону, Томасу Дж. Вирджину, Карлу Сторку, Джеральду Бернштейну и Джерами Дж. Сулливану.

Особо хочу сказать о группе замечательных коллег, которые сделали замечания для текста данного издания: Рональд Бремер, Техасский технологический университет; Стержиос Фотопоулос, Вашингтонский государственный университет; Мишель Ганен, Уэбстерский университет; Филипп Муса, Техасский технологический университет; Томас Обремски, Денверский университет; Даррелл

Радсон, Висконсинский университет, штат Милуоки; Терренс Релли, Бэдсонский колледж; Питер Шухман, Ричмондский университет; Бала Шетти, Техасский университет, Л. Дуайт Снизен-мл., Аризонский университет; Тед Тзукахара, колледж Святой Марии; Эдвард А. Усил, Американский университет; Мишель Уэгманн, школа менеджмента Келлера; Мустафа Йилмаз, Северо-восточный университет; и Гейри Йошимото, государственный университет Святого Клода.

Мои слова признательности также всем тем, чей вклад в предыдущие издания помог принять этой книге завершенный вид: Сангиту Чаттерджи, Северо-восточный университет; Джей Девоур, Калифорнийский государственный политехнический университет; Берту Холланду, Темплийский университет; Уинстону Лину, государственный университет Нью-Йорка в Буффало; Герберту Спиреру, Коннектикутский университет; Дональду Уэстерфильду, Уэбстерский университет; Уэйну Уинстону, Индианский университет; Джеку Юркевичу, Пейсовский университет; Бетти Торн, Стетсоновский университет; Деннису Петруска, Янгстаунский государственный университет; Х. Кариму, университет Западного побережья; Мартину Янгу, Мичиганский университет; Ричарду Спинетто, университет Колорадо в Булстере; Полю Пашке, Орегонский государственный университет; Ларри Амманну, Техасский университет в Далласе; Дональду Марксу, университет Аляски; Кевину Нгу, университет Оттавы; Рахмату Тавалляли, Уэльский университет; Дэвиду Ауэру, Западно-Вашингтонский университет; Мюрею Коте, Техасский университет; Петеру Лакнеру, Нью-йоркский университет; Дональду Адольфсону, университет Бригхам Янга; А. Рахулджи Парса, Дрейковский университет.

Обращение к студентам

Когда вы начинаете этот курс, у вас может быть некоторое предвзятое мнение о том, что представляет собой статистика. Если ваше мнение положительное, сохраните его и поделитесь им с однокурсниками. Если же негативное — то отложите его в сторону и оставайтесь непредубежденным. Поскольку этот предмет часто не изучают в высших учебных заведениях многих стран, то относительно небольшое количество студентов развили в себе статистический взгляд на мир, взгляд, который помогает им справляться с горами данных при принятии решений в условиях неопределенности.

По нескольким причинам статистика легче для вашего поколения, чем для предыдущих. Сейчас, когда компьютеры могут делать за вас черновую работу по проведению расчетов данных, у вас есть время глубже изучить основные понятия статистики и осознать, как это поможет вам победить конкурентов в бизнесе.

Хорошо усвойте вводный материал, чтобы понять, почему статистику стоит изучать. Пользуйтесь расположенными в конце глав резюме, списками основных понятий и другими материалами. Не забывайте, что в конце книги приведены подробные решения задач, а также словарь терминов, который всегда поможет вам быстро вспомнить нужное слово. И не волнуйтесь. Если вы осознали необходимость статистики для вашего бизнеса, вещи, которые вы должны выучить, легко найдут свое место.

Храните книгу как справочник. Она вам пригодится, когда ваш босс попросит немедленно написать докладную записку, которая требует быстрого просмотра данных, или подготовить ответ на аналитический материал конкурентов. С помощью *Практической бизнес-статистики, которая будет стоять у вас в шкафу*, вы быстро закончите работу и пойдете на обед. *Приятного аппетита*!

Эндрю Ф. Сигел (Andrew F. Siegel)

Об авторе

Эндрю Ф. Сигел (Andrew F. Siegel) — профессор факультета менеджмента и финансов школы бизнеса университета штата Вашингтон, Сиэтл. Он также является адъюнкт-профессором факультета статистики и факультета молекулярной биотехнологии, имеет звание доктора философии по статистике Станфордского университета (1977 г.), магистра наук по математике Станфордского университета (1975 г.) и бакалавра по математике и физике (с отличием) Бостонского университета (1973 г.). До работы в Сиэтле он преподавал и занимался исследованиями в Гарвардском университете, университете штата Висконсин, в корпорации RAND, Смитсоновском институте и в Принстонском университете. Он также периодически читал лекции (как приглашенный профессор) в Бургундском университете в Дижоне и в Сорбонне, Франция. Впервые преподавая статистику в школе бизнеса (университет штата Вашингтон, 1983 г.), он был удостоен звания лучшего профессора семестра на основании опроса студентов MBA. В 1993 году он получил должность профессора, которая финансируется грантом Батербау; эта профессорская должность была учреждена в честь выдающегося профессора И. Батербау (I. Butterbaugh), преподавателя бизнес-статистики. Другие его награды и премии: премии Бурлингтонского Северного Фонда за выдающиеся достижения (1986 и 1992 гг.); член-корреспондент Центра по изучению фьючерских рынков, Колумбийский университет, 1988 г.; награды за успехи в преподавании (исполнительная программа MBA), университет штата Вашингтон, 1986 и 1988 гг.; награда Фонда Пита Марвика (Peat Marwick Foundation) за исследование возможностей аудита, 1987 г., и Phi Beta Kappa, 1973 г.

Эндрю Ф. Сигел является членом Американской статистической ассоциации (American Statistical Association), где занимал должность секретаря-казначея секции экономической и бизнес-статистики. Им написаны также книги *Статистика и анализ данных. Введение* (*Statistics and Data Analysis: An Introduction*, Wiley, 1996, совместно с Charles J. Morgan), *Контрпримеры в вероятности и статистике* (*Counterexamples in Probability and Statistics*, Wadsworth, 1986, совместно с Joseph P. Romano) и *Современный анализ данных* (*Modern Data Analysis*, Academic Press, 1982, совместное редактирование с Robert L. Launer). Его статьи опубликованы в следующих изданиях: *Journal of the American Statistical Association, Journal of Business, Management Science, Journal of Finance, Encyclopedia of Statistical Sciences, American Statistician, Journal Financial and Quantitative Analysis, American Mathematical Monthly, Journal of the Royal Statistical Society, Annals of Statistic, Annals of Probability, Society for Industrial and Applied Mathematics Journal on Scientific and Statistical Computing; Journal of*

Computational Biology, Genome Research, Biometrika, Auditing: A Journal of Practice and Theory, Contemporary Accounting Research; Journal of Futures Markets и Journal of Applied Probability. Он работал консультантом в различных прикладных областях, таких как прогнозирование результатов выборов для крупной телевизионной сети, статистические алгоритмы распознавания речи для известной исследовательской лаборатории, тестирование телевизионной рекламы для маркетинговой фирмы, методы контроля качества продукции поставщиков для крупной промышленной компании, эффективность биотехнологических процессов в крупной лаборатории, автоматизация проектирования и запуск в производство электронного оборудования в Силиконовой Долине, анализ диверсификации портфеля активов финансовой компании.

ЧАСТЬ I

Введение и описательная статистика

В этой части...

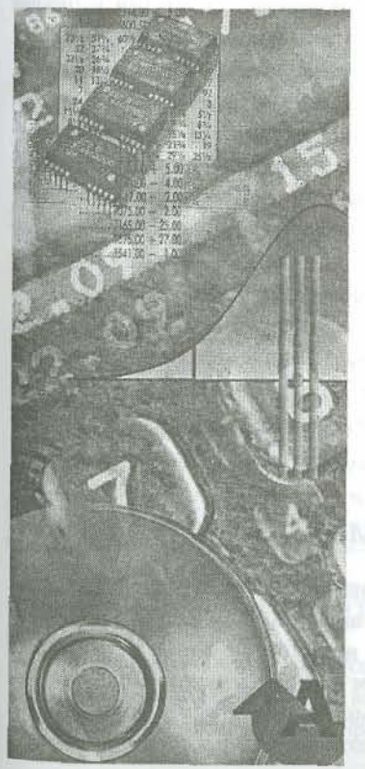
Глава 1. "Введение: роль статистики в бизнесе"

Глава 2. "Структуры данных: классификация различных типов наборов данных"

Глава 3. "Гистограммы: взгляд на распределение данных"

Глава 4. "Обобщающие показатели: интерпретация типических значений и перцентилей"

Глава 5. "Изменчивость: изучение разнообразия"

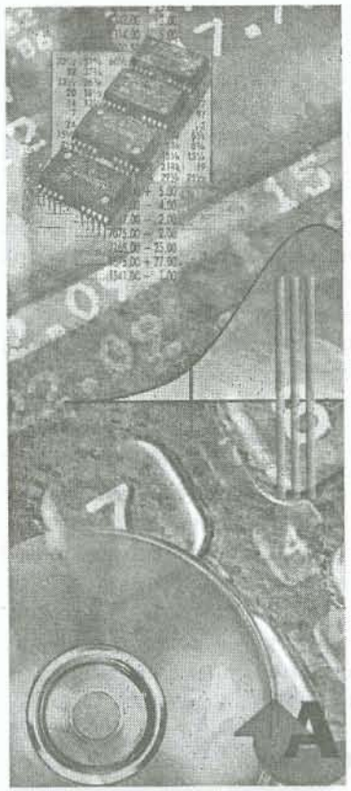


Добро пожаловать в мир статистики! Этот мир необходим вам, поскольку, когда вы располагаете информацией и знаете, как собрать необходимые дополнительные факты, вы сможете принять наилучшее управленческое решение. Как иначе можно руководить работой 12 отделов, 5809 служащими и производством 683 изделий? И даже для малого бизнеса необходимо понимать, что происходит в более широком бизнес-окружении, состоящем из потенциальных потребителей и конкурентов. Первые пять глав знакомят с ролью статистики в управлении экономической деятельностью (глава 1) и различными типами наборов данных (глава 2). Статистические показатели помогают увидеть "общую картину", которая иначе не проявилась бы в собранных данных. Глава 3 демонстрирует, как можно увидеть основные факты, касающиеся наборов чисел, с помощью *гистограмм*. Основные числовые показатели (среднее, медиана, перцентили и др.) представлены в главе 4. Одна из причин, по которой статистические методы имеют такое важное значение, заключается в том, что ситуации характеризуются *изменчивостью*, которая находит свое отражение в данных. В главе 5 рассказывается, как измерять степень такого разнообразия в данных.

ГЛАВА 1

Введение: роль статистики в бизнесе

Люди, занимающиеся бизнесом, вынуждены принимать решения в условиях постоянного давления обстоятельств, зачастую не имея полной и точной информации. Конечно же, любую доступную информацию следует использовать максимально полно. *Статистический анализ* помогает извлекать информацию из данных и оценивать качество этой информации. *Вероятность* помогает понять риски и случайности и обеспечивает оценки правдоподобности получения различных потенциальных результатов.



Даже те, кто считает, что принятие решения в бизнесе должно быть основано на интуиции и опыте специалистов (и, следовательно, не нуждается в излишней количественной информации), должны согласиться, что следует принимать во внимание всю имеющуюся ценную информацию. Таким образом, статистические методы следует рассматривать как важную часть процесса принятия решений, позволяющую выработать обоснованные стратегические решения, сочетающие интуицию специалиста с тщательным анализом имеющейся информации.

Мы начнем с обзора тех преимуществ, которые обеспечивает знание статистических методов, затем рассмотрим некоторые основные положения статистики и теории вероятности и их роль в экономической деятельности.

1.1. Почему именно статистика?

Действительно ли знание статистики так необходимо для успешного ведения бизнеса? Или достаточно полагаться на интуицию, опыт и догадку? Давайте сформулируем вопрос иначе: согласны ли вы игнорировать большую часть потенциально полезной информации, которая содержится в данных?

Почему необходимо изучать статистику?

Изучив статистику, вы станете более компетентно работать с данными и будете чувствовать себя гораздо увереннее в неопределенных ситуациях. Часто данные содержат много информации, которая не является очевидной, — статистика поможет извлечь и понять эту информацию. Нужна очень высокая квалификация, чтобы разрабатывать стратегию на основе знаний, опыта и интуиции. Если знания представлены в виде наборов чисел, статистика поможет ответить на такие вопросы. Насколько можно доверять этим числам и выводам из них? Как можно обобщить все эти данные? Используя статистику для получения знаний, вы приумножите свой опыт, что, безусловно, поможет принять правильное стратегическое решение.

Не следует пренебрегать статистикой. Методы статистики постоянно используются во всем мире, а снижение стоимости вычислительной техники увеличивает возможность принятия решений на основе количественной информации.

Сложна ли статистика?

Статистика не сложнее любой другой науки. Конечно, чтобы понять основные идеи и концепции, нужно хорошо поработать. Хотя некоторое внимание следует уделить отдельным деталям и вычислениям, все же намного легче стать опытным *пользователем* статистики, чем опытным статистиком, разбирающимся во всех мелочах. Сейчас стало легче использовать статистику, поскольку компьютеры теперь легко выполняют большое количество однотипных вычислений, что позволяет сконцентрироваться на содержательной интерпретации результатов. Возможно, некоторые непреклонные чистые статистики недовольны отсутствием изложения технических деталей при обучении статистическим методам, но сейчас уже хорошо видно, что эти детали заняли подобающее им место. Жизнь слишком коротка, чтобы человек посвящал ее разбору таких сложных технических приемов, как деление в столбик или обращение матрицы.

Влияет ли знание статистики на гибкость принятия решений?

Знание статистики помогает вам принимать хорошие решения. Статистика — не жесткая наука, она не отвергает опыт и интуицию. Получив знания о работе с данными и об основных свойствах неопределенных событий, вы улучшите информационную основу принятия решений и разовьете свою интуицию. Рассматривайте статистику как один из компонентов процесса принятия решений, но не как весь процесс. Статистика дополняет, но не заменяет деловой опыт, здравый смысл и интуицию.

1.2. Что такое статистика?

Статистика — это искусство и наука сбора и анализа данных. Поскольку *данными* называют любой вид зарегистрированной информации, статистика играет важную роль во всех сферах деятельности человека.

Статистика рассматривает общую картину

Если вы имеете большой и сложный набор данных, состоящий из многих небольших порций информации, статистика поможет классифицировать и проанализировать ситуацию, предоставив полезный обзор и резюме основных характеристик этих данных. Если данные пока отсутствуют, статистика поможет собрать их, обеспечивая получение ответов на ваши вопросы, а также то, что вы затратите не слишком большие усилия на выполнение этой работы.

Статистика не игнорирует отдельные объекты

При правильном подходе статистика уделяет необходимое внимание каждому из изучаемых объектов. Полный и тщательный статистический анализ обобщит основные свойства, которые характерны для каждого из объектов, а также *подготовит вас к интерпретации любых исключений*. Если данные содержат особые случаи, которые неадекватно отражены в "общей картине", то работа специалиста-статистика еще не завершена. Например, вы можете прочитать, что в 1996 году средний размер американской семьи составлял 2,65 человека[1]. Хотя это полезная статистическая информация, она не дает полную картину размера всех семей в США. Далее вы узнаете, что с помощью статистических методов можно легко описывать полное распределение размеров всех семей в США.

Пример. Данные в менеджменте

Данные очень часто используют в менеджменте. Ниже приведен краткий перечень видов ежедневно используемой в менеджменте информации (по сути, перечень используемых данных).

1. Финансовые отчеты (и другие виды бухгалтерской отчетности).
2. Курсы и объемы ценных бумаг, процентные ставки (и другая информация, относящаяся к инвестициям).
3. Состояние бюджета (и другие сообщения правительства).
4. Отчеты о продажах (и другие внутренние отчеты).
5. Результаты обзора состояния рынка (и другие маркетинговые отчеты).
6. Данные о качестве продукции (и другие производственные отчеты).
7. Отчеты о производительности рабочих (и другие внутренние данные фирмы).
8. Цена и объем проданной продукции (и другие данные о продажах).
9. Расходы на рекламу и результаты рекламной компании (и другая рекламная информация).

Подумайте об этом. Вероятно, большая часть вашей деятельности зависит, по крайней мере, косвенно от данных. Вероятно, кто-то работает для вас и консультирует вас по этим вопросам, но вы редко видите фактические данные. Время от времени вы могли бы попросить дать вам "сырые данные", чтобы иметь представление о перспективе. Просмотрев данные и задав вопросы, можно получить неожиданные результаты: можно обнаружить, что качество данных не так высоко, как вы думали (и у вас появляется мысль: на чем же мы строим наши прогнозы?), или, наоборот, вы получите обнадеживающие результаты и обретете уверенность. Другими словами, данные заслуживают внимания.

[1] *Statistical Abstracts of the United States*: 1997 (117 edition) Washington, DC, 1977, p. 59.

Посмотрим на данные

О чем вы думаете, глядя на таблицы с данными (например, на последние страницы *The Wall Street Journal*)? И что видит в них профессиональный статистик? Удивительно, но в обоих случаях ответ, как правило, один: не так уж и много. Нужно поработать с числами: нарисовать на бумаге графики, вычислить ряд характеристик и сделать еще много работы, прежде чем прояснится содержащаяся в этих числах информация. Именно этим занимаются профессиональные статистики. Они считают, что это проще и полезнее, чем подолгу подробно рассматривать длинные столбики чисел. Так что пусть вас не обескураживает, если столбик чисел выглядит для вас не более чем столбик чисел.

Статистика в менеджменте

Что должен знать о статистике менеджер? Вы должны иметь представление об основных понятиях статистики, включая некоторые (но не обязательно все) детали. Следует осознавать случайность и неопределенность многих аспектов окружающего нас мира. Более того, вы должны:

1. понимать и использовать в качестве базовой для вашей деятельности информации результаты статистического анализа;
2. принимать непосредственное участие в статистическом исследовании, если вы отвечаете за сбор и/или анализ данных.

Для этого не обязательно самому уметь выполнять сложный статистический анализ, но чтобы правильно интерпретировать полученные результаты, необходим некоторый опыт статистического анализа данных, который поможет вам также научить других глубже вникать в результаты. Более того, иногда удобнее выполнить определенный статистический анализ самостоятельно.

В последующих разделах мы уделим основное внимание идеям и понятиям статистики, сопровождая их для лучшего понимания практическими примерами.

1.3. Четыре основных этапа статистического анализа

На начальной стадии статистического анализа данные или не собраны, или еще даже не принято решение о том, какие данные следует изучать. Эти вопросы решают на этапе *планирования* таким образом, чтобы получить действительно полезную информацию. Когда данные есть в наличии, на этапе *исследования* проводится их первичный (предварительный) анализ. Следующий этап — *оценки* — позволяет получить на основе данных числовое значение неизвестной величины. Наконец, на последнем этапе — *проверке гипотез* — данные используются для принятия решения о соответствии выдвинутого предположения действительности. Рассмотрим все эти этапы по очереди.

Планирование сбора данных

Планирование сбора данных в маркетинговых исследованиях называют *планированием выборочного исследования*, а в изучении оптимизации химического производственного процесса — *планированием эксперимента*. Эта стадия пла-

нирования исследования включает составление подробного плана сбора данных. Тщательное составление плана поможет избежать лишних расходов и разочарования, если окажется (и будет уже слишком поздно), что собранные данные неадекватны основным поставленным вопросам. Разумный план также включает определение необходимого объема данных, достаточного для анализа, но не настолько большого, чтобы быть излишне расточительным. Таким образом, заранее составленный план удерживает стоимость проекта в разумных рамках и гарантирует, что стадия анализа будет протекать достаточно гладко.

Статистика особенно полезна тогда, когда есть большая группа людей, фирм или других объектов (*генеральная совокупность*), которая вас интересует, но вы не можете себе позволить провести полное исследование. Вместо этого, чтобы получить полезное, но неидеальное понимание ситуации в этой генеральной совокупности, можно отобрать небольшую группу (*выборку*), состоящую из некоторых (но не всех) объектов генеральной совокупности. Процесс обобщения результатов исследования выборки на всю совокупность называется *статистическим выводом*. *Случайная выборка* является одним из наилучших способов извлечь для подробного изучения выборку из генеральной совокупности, которая слишком велика, чтобы ее можно было изучать полностью[2]. Случайное извлечение выборки преследует две цели.

1. Гарантировать, что процесс извлечения выборки является беспристрастным, т.е. все объекты генеральной совокупности имеют равные шансы быть отобранными. Поэтому в среднем выборки являются представительными (репрезентативными) для данной генеральной совокупности (хотя каждая отдельная случайная выборка обычно является репрезентативной только приближенно, но не идеально).
2. Случайность, введенная контролируемым способом на стадии планирования проекта, гарантирует валидность (корректность) последующих статистических выводов.

Предварительное исследование данных

Как только вы получаете набор данных, вам хочется проверить его. В ходе **предварительного исследования** данные анализируются с разных точек зрения, описываются и обобщаются. Это позволяет убедиться, что данные представляют собой именно то, что необходимо, и нет никаких очевидных проблем[3]. Но хорошо выполненное предварительное исследование двояко готовит вас к проведению формального анализа.

1. Путем проверки, что ожидаемые связи действительно существуют в данных, и, таким образом, запланированные методы анализа адекватны данным.
2. Путем обнаружения в данных неожиданной структуры, которую необходимо принять во внимание, что предполагает внесение изменений в план анализа.

[2] Подробности построения случайной выборки изложены в главе 8.

[3] Предварительное исследование данных используется там, где это необходимо (и особенно в главах 3, 4, 11, 12 и 14).

Предварительное исследование является первой стадией. Часто недостаточно полагаться на формальный, автоматизированный анализ, который предполагает, что набор данных, вводимых в компьютер, "ведет себя хорошо". Всякий раз при возможности самостоятельно изучайте данные, чтобы убедиться, что все в порядке, т.е. нет больших ошибок и наблюдаемые в данных зависимости между параметрами соответствуют типу запланированного анализа. Эта стадия поможет внести в данные коррективы, выбрать соответствующий метод анализа и обосновать использование необходимых в дальнейшем статистических методов.

Оценка неизвестной величины

Оценка неизвестной величины представляет собой наиболее обоснованное, основанное на имеющихся данных, предположение о возможном значении. Поэтому желательно (а часто необходимо) оценивать те параметры, которые невозможно определить точно. Ниже приведено несколько примеров неизвестных величин для оценивания.

1. Объем продаж в следующем квартале.
2. Намерения правительства по изменению налоговых ставок.
3. Реакция населения Сиэтла на новый продукт.
4. Стоимость портфеля инвестиций в следующем году.
5. Изменение производительности при изменении стратегии.
6. Уровень брака в производственном процессе.
7. Победители следующих выборов.
8. Влияние продолжительного воздействия излучения экрана компьютера на здоровье.

Статистика может пролить свет на некоторые из этих ситуаций, предоставив хорошо обоснованное предположение исходя из надежных данных. Помните, что все статистические оценки являются только предположениями, а следовательно, часто бывают неточны. Однако они служат поставленным целям, если достаточно близки к соответствующим неизвестным величинам. Если известно, насколько (приближенно) точны эти оценки, то можно решить, в какой мере их стоит принимать во внимание.

Статистическая оценка также показывает величину неопределенности или ошибки в некотором предположении, рассчитанном для выборки, случайно взятой из большей по размеру генеральной совокупности. *Доверительный интервал* дает вероятное значение верхней и нижней границ оцениваемой неизвестной величины, что позволяет заявить: "Я не знаю точное значение неизвестной величины, но я достаточно уверен в том, что оно лежит между этими двумя числами".

Обычно вычисляют доверительные интервалы, поскольку они показывают, насколько надежной в действительности является оценка. Например, утверждение, что в следующем квартале объем продаж составит $11 300 000, содержит некоторую определенную информацию. Однако утверждение, что вы на 95% уверены в том, что в следующем квартале объем продаж будет находиться в пределах от $5 900 000 до $16 700 000, позволяет делать дополнительные и гораздо более глубокие выводы.

Доверительные интервалы представляют оценку в некоторой перспективе и позволяют избежать необходимости указывать одно число как точное значение, в то время как фактически это число точным не является[4].

Проверка гипотез

Проверка статистических гипотез заключается в использовании данных для осуществления выбора одной из двух (или более) различных возможностей при решении вопроса в неоднозначной ситуации. Проверка гипотезы исходя из данных дает определенное решение о том, какая из возможностей является верной. Процедура проверки гипотезы включает сбор данных, которые помогают осуществить выбор одной из возможностей, и использование статистического анализа для подтверждения принятого решения, если это решение не вытекает из простого беглого анализа данных[5].

Ниже приведено несколько примеров гипотез, которые можно было бы проверить с использованием данных.

1. Средний житель Нью-Йорка в следующем месяце планирует тратить на покупку вашего продукта по крайней мере $10.
2. Завтра на выборах вы победите.
3. Новое медицинское средство безопасно и эффективно.
4. Средство марки "X" эффективно стирает и отбеливает.
5. Ошибка в финансовом отчете меньше некоторой величины.
6. Исходя из прошлого опыта можно предсказать ситуацию на фондовом рынке.
7. Уровень производственного брака ниже, чем ожидают потребители продукции.

Обратите внимание, что каждая гипотеза сформулирована как определенное утверждение, которое может быть либо верным, либо неверным. Результатом проверки гипотезы является заключение о том, что данные либо подтверждают гипотезу, либо нет.

Часто статистические методы используют, чтобы решить, можно ли в качестве допустимой возможности рассматривать просто "чистую случайность". Например, если опрос 300 человек свидетельствует, что 53% планируют завтра голосовать за вас, можно ли сделать вывод, что выборы закончатся в вашу пользу? Несмотря на то что здесь можно поставить много вопросов, давайте сейчас отвлечемся от деталей, как, например, от реальной возможности того, что некоторые люди до завтра изменят свое мнение, и сконцентрируемся только на элементе случайности (обусловленной тем фактом, что невозможно опросить всех избирателей и узнать предпочтение каждого из них). В этом примере тщательный анализ покажет, что существует реальная возможность того, что меньше 50% избирателей предпочтут вас, а наблюдение значения 53% находится в пределах диапазона ожидаемого случайного изменения.

[4] Подробно доверительные интервалы изложены в главе 9 и используются в главах 9–15.

[5] Подробно процедура проверки статистических гипотез изложена в главе 10 и используется в главах 10–18.

Пример. *Статистический контроль качества*

Ваши производственные процессы несовершенны (как и у других фирм), и время от времени некоторое изделие необходимо или вторично переработать, или просто выбросить. Нужно сказать спасибо вашей группе контроля, которая делает все, чтобы недоброкачественное изделие не попало к потребителю. Однако контроль, обнаружение и решение этих проблем — все это стоит немалых денег. Вот почему многие фирмы начали использовать статистический контроль качества.

Упрощая ситуацию, будем считать, что ваша сборочная линия *контролируема*, если изготовленные изделия имеют стабильные показатели, которые удовлетворяют техническим требованиям. В противном случае, сборочная линия считается *неконтролируемой*. Статистические методы помогают наблюдать за производственным процессом таким образом, что можно сэкономить финансовые средства тремя способами: (1) снизить затраты на сбор информации; (2) быстро выявлять проблемы и, следовательно, минимизировать ущерб; (3) по возможности не вмешиваться в процесс тогда, когда в этом нет необходимости. Ниже в общих чертах описано, как в данной ситуации реализуются четыре этапа статистического анализа.

На стадии планирования следует решить, *что* и *как часто* следует измерять. Например, можно принять решение извлекать случайную выборку объемом 5 изделий из каждой партии объемом 500 изделий. Каждое изделие в выборке оценивают по внешнему виду, выявляя очевидный брак, а также измеряют длину и ширину изделия. Результат стадии планирования исследования представляет собой план раннего выявления проблем. План должен работать в *реальном времени*, чтобы проблемы можно было выявлять немедленно, а не на следующей неделе.

В ходе предварительного исследования данные наносят на *карты контроля качества* и изучают те конфигурации, которые вызывают тревогу. Правильно определив направление изменения данных, можно даже предсказать и установить проблему раньше, чем она приведет к производственным потерям!

Статистическая оценка обеспечивает менеджмент информацией о ходе производственного процесса. Если производственный процесс хорошо управляется в установленных границах, то можно даже поднять сортность продукции, а значит, и цену. Оценки качества текущей продукции необходимы для удовлетворения текущих заказов, а прогноз качества на будущее полезен для стратегического планирования и выработки ценовой политики.

Статистическую проверку гипотез можно использовать для ответа на важный вопрос: контролируется данный процесс или нет? Поскольку производственный процесс может быть большим, длительным и сложным, не всегда можно оценить его, посмотрев на работу части оборудования. Максимально используя статистическую информацию, содержащуюся в имеющихся данных, вы надеетесь достичь двух целей. Во-первых, вы хотите определить момент выхода системы из-под контроля, прежде чем уровень качества станет недопустимым. Во-вторых, вам хочется минимизировать "ложную тревогу", чтобы не тратить напрасно время и деньги на вмешательство в процесс тогда, когда он фактически является управляемым.

Пример. *Запуск нового продукта*

Решение вопроса об освоении производства нового продукта является одним из самых важных, принимаемых компанией, и для этого компании необходимо располагать большим количеством информации. Большую часть этой информации получают в результате статистических исследований. Например, маркетинговое исследование целевой группы потребителей можно было бы использовать, чтобы оценить, какое количество людей хотели бы приобрести новое изделие по каждой из предложенных цен. Прошлые данные о себестоимости аналогичных типов изделий можно было бы использовать для оценки себестоимости нового продукта. Анализ прошлого опыта освоения выпуска изделий, как удачный, так и неудачный, можно использовать как руководство при запуске нового продукта. Анализ статистических данных о национальных и международных фирмах, выпускающих аналогичный продукт, поможет вам оценить масштаб будущей конкуренции. Прежде чем вкладывать финансовые средства в несколько отобранных рекламных роликов, неплохо было бы проверить реакцию выборки возможных потребителей на отдельные виды рекламы.

Четыре основных этапа статистического анализа могут в такой ситуации быть реализованы различными способами. Поскольку генеральная совокупность потребителей слишком велика, чтобы можно было изучить ее полностью, можно *запланировать* исследование репрезентативной выборки из генеральной совокупности (например, чтобы определить количество потребителей, готовых приобрести новый продукт, или чтобы посмотреть реакцию потребителей на конкретный рекламный ролик). Если есть данные, их всегда можно подвергнуть *предварительному исследованию*, что позволит получить более ясное представление о ситуации (например, можно ли в предположении о сегментации рынка выделить определенные группы потребителей?), а кроме того, выполнить обычную проверку данных перед использованием других статистических процедур. Можно вычислять различные оценки, которые, например, показывают потенциальный размер рынка, возможный первоначальный уровень продаж и себестоимость продукции. И, наконец, можно *проверять различные гипотезы*, чтобы, например, подтвердить гипотезу о достаточно высоком интересе потребителя и таким образом оправдать продолжение проекта или чтобы проверить эффективность рекламных роликов и выбрать тот из них, который (не случайно) лучше других с точки зрения реакции потребителей.

1.4. Что такое вероятность

Вероятность — это средство для работы с риском и неопределенностью. **Вероятность** показывает возможность (или шанс) наступления в будущем каждого из различных потенциальных событий, рассчитанную на основании информации о некоторой ситуации. Например, можно допустить, что вы располагаете почти всей информацией об интересующей вас ситуации (т.е. известны все подробности осуществления процесса, который приводит к успеху, неудаче или некоторым потерям). Тогда можно вычислить вероятность получения различных результатов для разных стратегий, чтобы увидеть преимущество каждой из них.

Например, вы знаете, что шансы на успех международного проекта равны только 8% (т.е. вероятность равна 0,08), но вы предполагаете, что правительство сможет удержать инфляцию на низком уровне, и тогда шанс на успех возрастет до 35% (риск все еще высок, но все же лучше, чем 8%). Вероятность не подскажет вам, стоит ли инвестировать проект, но зато вы сможете трезво оценить ситуацию.

Ниже приведены примеры различных ситуаций, где для принятия решения необходимо вычислить или оценить значение вероятности.

1. При заданной структуре инвестиционного портфеля и определенных предположениях о работе финансовых рынков, чему равна вероятность получения прибыли в течение одного года? В течение 10 лет?
2. Насколько велика вероятность того, что завтра пойдет дождь? Чему равна вероятность того, что наступающая зима будет достаточно холодной, чтобы ваш бизнес, связанный с масляными радиаторами, был прибыльным?
3. Чему равна вероятность того, что иностранная держава (в которой находится ваш завод) будет втянута в гражданскую войну в течение двух следующих лет?
4. Чему равна вероятность того, что студент колледжа, прошедший анкетирование при устройстве на работу, в течение ближайших месяцев станет ценным работником?

Вероятность — это понятие, в некотором смысле обратное статистике. В то время как статистика помогает переходить от наблюдений к обобщениям относительно рассматриваемой ситуации, вероятность имеет обратную направленность — исходя из характеристики ситуации можно выяснить, какие данные вы скорее всего получите, и возможность получения каждого из вариантов таких данных. Эта обратная связь может быть графически изображена следующим образом.

$$\left(\begin{array}{l}\text{Характеристика}\\ \text{ситуации}\end{array}\right) \xrightarrow{\text{Вероятность}} \left(\begin{array}{l}\text{Какое событие вероятнее}\\ \text{всего произойдет}\end{array}\right).$$

$$(\text{Какое событие произошло}) \xrightarrow{\text{Статистический вывод}} (\text{Характеристика ситуации}).$$

Вероятность всегда идет рядом со статистикой, обеспечивая прочный фундамент для статистического вывода. В условиях неопределенности нельзя знать точно, какое событие произойдет, всегда есть некоторая вероятность ошибки. Используя понятие вероятности, вы узнаете, как контролировать ошибку так, чтобы она наблюдалась не более чем в 5% или 1% случаев.

1.5. Общий совет

Статистические результаты должны допускать простое непосредственное объяснение (даже если соответствующая теория намного сложнее). Ниже приведено несколько общих советов.

1. Доверяйте своим суждениям, учитывайте здравый смысл.
2. Сохраняйте здоровый скептицизм.
3. Не дайте себя ввести в заблуждение с помощью на первый взгляд оригинального статистического анализа. Он может опираться на нереальные или неподходящие предположения.

Из-за большой гибкости, доступной аналитику на каждой стадии статистического анализа, один из самых важных факторов, который надо принять во внимание при оценке результатов статистического анализа, звучит так: *Кто это финансировал*? Помните, что аналитик много раз делает выбор — при определении проблемы, при планировании сбора данных, при выборе структуры или модели для анализа, при интерпретации результатов.

1.6. Дополнительный материал

Резюме

Статистика — это искусство и наука сбора и анализа данных. Статистические методы следует рассматривать как важную часть процесса принятия решений, позволяющую вырабатывать обоснованные стратегические решения, сочетающие интуицию специалиста с тщательным анализом имеющейся информации. Использование статистики становится все более важным преимуществом в конкуренции.

Ниже приведены основные этапы статистического анализа.

1. **Планирование исследования** включает составление подробного плана сбора данных, возможно, с использованием случайной выборки из генеральной совокупности.
2. **Предварительное исследование данных** включает рассмотрение набора данных с разных точек зрения, описание и обобщение данных. Выполнение этого этапа помогает убедиться, что запланированный анализ адекватен данным, а при необходимости позволяет внести в процесс анализа определенные коррективы.
3. **Оценивание неизвестной величины** дает наиболее обоснованное возможное предположение о значении, основанное на исходных данных. Кроме того, есть возможность вычислить величину ошибки, которая возникает при использовании оценки вместо фактического, но неизвестного значения.
4. **Проверка статистических гипотез** заключается в использовании данных для выбора одной из двух (или больше) различных возможностей при решении вопроса в неопределенной ситуации. Такая проверка позволяет убедиться, действительно ли данные обладают определенным интересным свойством, или мы имеем дело с "чистой случайностью", которая не представляет интереса.

Вероятность исходя из предположений об изучаемой ситуации показывает возможность или шанс наступления в будущем каждого из нескольких потенциальных событий. Вероятность — это понятие, в некотором смысле обратное статистике: вероятность показывает, какие данные вы скорее всего получите, если известна характеристика ситуации, а статистика помогает охарактеризовать ситуацию в результате анализа и обобщения данных.

Статистика лучше всего работает в сочетании с вашими собственными экспертными заключениями и здравым смыслом. Если результаты статистического анализа расходятся с вашей интуицией, необходимо разобраться, чтобы установить причину. Статистический анализ может оказаться некорректным, если в его основу положены неверные допущения, или ваша интуиция может вас подвести, если она не основана на фактах.

Основные термины

- Статистика (statistics), 29
- Планирование исследования (designing the study), 31
- Предварительное исследование данных (exploring the data), 32
- Оценка неизвестной величины (estimating an unknown quantity), 33
- Проверка статистических гипотез (hypothesis testing), 34
- Вероятность (probability), 36

Контрольные вопросы

1. Почему стоит тратить усилия на изучение статистики?

   а) Дайте ответ для менеджмента в целом.

   б) Дайте ответ для конкретной, интересующей вас сферы экономической деятельности.
2. Выберите коммерческую фирму и укажите, как можно использовать статистический анализ для принятия решений в отношении деятельности этой фирмы.
3. Как сочетаются между собой статистический анализ и экономический опыт?
4. Что такое статистика?
5. Что представляет собой стадия планирования в статистическом исследовании?
6. Почему лучшим методом отбора объектов для анализа является случайный отбор?
7. В чем польза предварительного исследования данных в дополнение к результатам автоматического компьютерного анализа?
8. Всегда ли корректны статистические оценки? Если нет, то что еще (в дополнение к оценкам значений) необходимо иметь для более эффективного их использования?
9. Почему доверительный интервал полезнее, чем оценка значения?
10. Приведите два примера проверки гипотезы о ситуации, которая могла бы заинтересовать коммерческую фирму.
11. В чем разница между вероятностью и статистикой?
12. Консультант представил очень сложный статистический анализ с большим количеством математических символов и уравнений. Результаты его анализа противоречат вашему опыту и интуиции. Что следует предпринять в данной ситуации?
13. Почему важно знать источник финансирования исследования при оценке результатов статистического анализа?

Задачи

1. Опишите последнее принятое вами решение, которое частично зависело от информации, полученной из данных. Определите этот набор данных и укажите, как знание статистики поможет вам в использовании этих данных.
2. Назовите три параметра, которые интересуют фирму, но для которых не известны точные числовые значения. Для каждого параметра укажите оценку, которая может оказаться полезной. Изложите в общих терминах, насколько надежными могут быть эти оценки.
3. Рассмотрите еще раз значения трех оценок из предыдущей задачи. Имеются ли для них доверительные интервалы? Если да, то насколько они могут быть полезны?

4. Назовите два вида обычно принимаемых вами решений, для которых может оказаться полезной проверка статистических гипотез.
5. Просмотрите последний номер *The Wall Street Journal*. Выделите статью, которая прямо или косвенно опирается на статистику. Кратко изложите содержание статьи (не забудьте указать название, дату и номер страницы). Какой из четырех этапов статистического анализа здесь представлен?
6. Какой из четырех этапов статистического анализа представлен в каждой из следующих ситуаций?

   а) Отдел контроля качества предприятия изучает подробную количественную информацию о текущей производительности, чтобы выявить возможные проблемы.

   б) Фокус-группа обсуждает целевую рекламу, чтобы подготовить соответствующий вопросник для исследования.

   в) Фирме предъявлено обвинение в дискриминации сотрудников по признаку пола. Присяжным представлены данные о заработной плате мужчин и женщин, чтобы убедить их в наличии дискриминации и в том, что имеющиеся различия не могут быть объяснены только лишь случайностью.

   г) Для прогнозирования объема продаж фирмы необходимо знать объем валового национального продукта в следующем квартале. Поскольку этих данных в настоящее время нет, используют некоторое обоснованное предположение.
7. За последний месяц по неизвестной причине резко упал объем продаж за рубежом. Более того, вы понимаете, что у вас даже нет цифр для постановки проблемы. Вы собираете совещание, чтобы выяснить этот вопрос. С каким из этапов статистического анализа вы будете иметь дело?
8. Если ваше предприятие производит слишком много продукции, то вам приходится платить за хранение излишков материальных запасов. Если вы производите слишком мало, то теряете потребителей и прибыль. Следовательно, желательно производить действительно необходимое количество продукции, чтобы избежать подобных потерь для вашей фирмы. Однако, к сожалению, вы не знаете, сколько точно необходимо производить продукции. Какой этап статистического анализа подходит для решения этой проблемы?
9. До начала анализа только что собранного большого набора бухгалтерских данных ваш начальник попросил внимательно просмотреть данные, чтобы выявить проблемы и неожиданности, а также гарантировать целостность данных. Определите, какой этап статистического анализа вы при этом будете выполнять.
10. Ваша рабочая группа хотела бы оценить размер рынка комплектующих для высококачественных стереофонических устройств в Новом Орлеане, но не может найти легко доступные и надежные данные. Какой этап статистической деятельности возникает на первоначальной стадии этого проекта?

11. Вам необходимо принять решение о том, кого и в каком количестве следует опросить и как обработать результаты, чтобы стоимость опроса была минимальной. Определите этап статистического анализа, реализованный в этой ситуации.
12. Вашей фирме предъявлено обвинение в дискриминации. Ваша защита оспаривает обвинение, исходя из того, что различие так мало, что его можно объяснить случайностью и что фактически дискриминации нет. Какой этап статистического анализа здесь реализуется?
13. Внимательно просмотрев графические данные, отдел маркетинга определил три отчетливых рыночных сегмента, различающихся своими потребностями и уровнем цен. С помощью реализации какого этапа статистического анализа получена эта полезная информация?
14. На основе выборочного исследования вы пытаетесь определить качество последней полученной партии строительных материалов. Какой этап статистического анализа вам в этом поможет?
15. Вы полагаете, что один из станков поломан, но вы не уверены, поскольку даже при нормальной работе этого станка небольшая часть изготовленных деталей была некачественной. Какой этап статистического анализа вы реализуете, принимая решение о том, действительно ли увеличилось количество бракованных деталей?

Проект

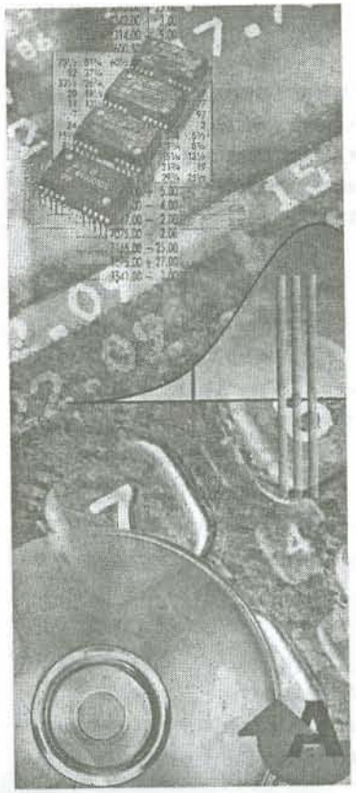


Найдите в газете, журнале или в Internet результаты опроса общественного мнения. Обсудите письменно, какой из четырех этапов статистического анализа был реализован при обработке результатов (если это видно из статьи) или мог быть реализован для выяснения мнения людей. Приложите к вашему обсуждению копию статьи.

ГЛАВА 2

Структуры данных: классификация различных типов наборов данных

Данные могут быть представлены в различной форме. Полезно иметь базовую классификацию различных типов данных, чтобы сразу же определять тип новых данных и использовать соответствующий метод анализа. **Набор данных** состоит из результатов наблюдений объектов, обычно включающих регистрацию одной и той же информации для каждого объекта. Мы определяем **элементарные единицы** как сами объекты (например, компании, люди, домохозяйства, города, телевизоры), чтобы отличать их от результатов измерений или наблюдений (например, объемы продаж, вес, доход, население, размер). Можно указать четыре основных способа классификации наборов данных.



Первый. По количеству порций информации (переменных) для каждой элементарной единицы.

Второй. По типу измерения (числа или категории) для каждого наблюдения.

Третий. По тому, важна или нет упорядоченность во времени записей о результатах измерений.

Четвертый. По тому, собиралась ли информация специально для этого анализа или данные собирались ранее кем-то другим для своих нужд.

2.1. Сколько переменных?

Порция информации, регистрируемая для каждого объекта (например, стоимость), называется **переменной**. Количество переменных, или порций информации, регистрируемых для каждого объекта, указывает на сложность набора данных и определяет соответствующий тип анализа. В зависимости от того, имеем ли мы дело с одной, двумя или многими переменными, мы получаем соответственно *одномерный*, *двумерный* или *многомерный* набор данных.

Одномерные данные

Одномерные наборы данных (одна переменная) содержат только один признак, зарегистрированный для каждой элементарной единицы. В этом случае статистические методы используют для обобщения основных свойств этого единственного признака, отвечая на такие вопросы.

1. Чему равно типичное (обобщенное) значение?
2. Насколько различаются эти объекты?
3. Имеются ли в этом наборе данных отдельные элементы или группы элементов, требующие особого внимания?

Указанная ниже таблица одномерных данных содержит объемы прибыли 12 компаний сферы общественного питания (из списка Fortune 500).

Компания	Прибыль в 1997 г., млн дол.	Компания	Прибыль в 1997 г., млн дол.
Advantica	–134,5	Outblack Steakhouse	61,5
Brinker International	60,5	Performance Food Group	13,2
Darden Restaurants	–91,0	ProSource	–13,7
Foodmaker	34,1	Shoney's	–35,7
Host Marriott Services	20,8	Viad	89,3
McDonald's	1 642,5	Wendy's International	130,5

Источник данных: `http:/www.pathfinder.com/fortune/fortune500/ind147.html`; осень 1998 г.



Приведем еще несколько примеров одномерных наборов данных.

1. Доходы отдельных людей, выявленные в ходе маркетингового исследования. Статистический анализ выявил бы структуру (или распределение) доходов, выявив типический уровень дохода, степень вариации доходов и процент людей, доход которых находится в любом заданном диапазоне.
2. Количество дефектов в каждом телевизоре из выборки объемом 50, взятой из телевизоров, изготовленных сегодня утром. Статистический анализ можно использовать для учета качества (оценивание) и наблюдения за тем, чтобы производственный процесс не вышел из под контроля (проверка гипотез).
3. Сделанные 25 экспертами прогнозы уровня процентной ставки. Анализ показал бы, как вы и подозревали, что оценки экспертов не согласуются и (если проверить позже) все они не верны. Хотя статистика не может на основе этих 25 оценок дать один точный прогноз, она, по крайней мере, позволит изучить данные на степень их согласованности.

4. Цвета, выбранные членами фокус-группы. Анализ поможет сделать подходящий выбор для нового вида продукции.
5. Оценки платежеспособности фирм в инвестиционном портфеле. Анализ показал бы риск инвестиционного портфеля.

Двумерные данные

Наборы **двумерных** (две переменные) данных содержат информацию о двух признаках для каждого из объектов. В дополнение к обобщению свойств каждой из этих двух переменных, рассматриваемых как отдельные наборы одномерных данных, статистические методы можно использовать для изучения связи между этими двумя измеренными факторами, выясняя при этом следующее.

1. Существует ли между этими двумя переменными простая связь?
2. Насколько сильно взаимосвязаны переменные?
3. Можно ли предсказать значение одной переменной на основании другой? Если да, то с какой степенью надежности?
4. Существуют ли отдельные объекты или группы, которые требуют особого внимания?

Приведенная ниже таблица содержит двумерные данные о размерах прибыли 12 компаний общественного питания (из списка Fortune 500) и изменения прибыли (в процентах) по сравнению с предыдущим годом.

Компания	Прибыль в 1997 г., млн дол.	Изменение прибыли, в % к 1996 году
Advantica	–134,5	0,0%
Brinker International	60,5	76,0
Darden Restaurants	–91,0	–222,4
Foobmaker	34,1	69,8
Host Marriott Services	20,8	45,5
McDonald's	1 642,5	4,4
Outback Steakhouse	61,5	–14,2
Performance Food Group	13,2	20,6
ProSource	–13,7	0,0
Shoney's	–35,7	–173,6
Viad	89,3	214,8
Wendy's International	130,5	–16,3

Источник данных: `http:/www.pathfinder.com/fortune/fortune500/ind147.html`; осень 1998 г.



Рассмотрим еще несколько примеров двумерных наборов данных.

1. Данные за прошлый квартал о затратах на производство продукции (первая переменная) и количестве произведенных изделий (вторая переменная) для каждой из семи фабрик (объекты или элементарные единицы), выпускающих интегральные схемы. Двумерный статистический анализ показал бы взаимосвязь между затратами и количеством произведенных интегральных схем. В частности, анализ определил бы *постоянные* затраты, связанные с использованием производственного оборудования, и *переменные* затраты, характеризующие производство одной дополнительной интегральной схемы[1]. Аналитик, посмотрев на данные о семи фабриках, мог бы сравнить их эффективность.
2. Цена одной обыкновенной акции вашей фирмы (первая переменная) и дата (вторая переменная), зарегистрированные каждый день на протяжении последних шести месяцев. Связь между ценой и временем позволяет увидеть тенденции в изменении стоимости инвестиций. Однако трудно сказать, можно ли на основании таких данных предсказать будущую стоимость инвестиций (это, в частности, зависит от того, является ли изменение стоимости непредсказуемым "случайным блужданием", или существует некоторая реальная закономерность).
3. Данные опроса 100 человек в торговом центре: приобреталось или не приобреталось некоторое изделие (первая переменная, записывается ответ "да/нет", или 1/0), и возможность вспомнить рекламу этого изделия (вторая переменная, записана аналогичным образом). Такие данные (и, конечно же, данные более подробных исследований) помогают пролить свет на эффективность рекламы, т.е. изучить связь между рекламой и покупкой.

Многомерные данные

Наборы **многомерных** (много переменных) данных содержат информацию о трех или более признаках для каждого объекта. В дополнение к обобщению свойств каждой из этих переменных (рассматриваемых как отдельные наборы одномерных данных) и установлению зависимости между парами переменных (как при анализе набора двумерных данных) статистические методы можно использовать для изучения взаимосвязей между всеми этими переменными, выясняя при этом следующие вопросы.

1. Существует ли простая зависимость между этими признаками?
2. Насколько сильно они взаимосвязаны?
3. Можно ли предсказать значения одной ("выделенной") переменной исходя из значений остальных? С какой степенью надежности?
4. Существуют ли отдельные объекты или группы, которые требуют особого внимания?

В представленной ниже таблице содержатся многомерные данные о размерах прибыли 12 компаний общественного питания (из списка Fortune 500) вместе с процентом изменения прибыли по отношению к предыдущему году, количеством служащих и размерами дохода.

[1] Переменной стоимостью называется стоимость, которая зависит от количества произведенных единиц продукции; это не связано с понятием *статистической* переменной.

Компания	Прибыль в 1997 г., млн дол.	Изменение прибыли, в % к 1996 г.	Количество служащих	Доходы, млн дол.
Advantica	–134,5	0,0	85 000	2 609, 5
Brinker International	60,5	76,0	47 000	1 335,3
Darden Restaurants	–91,0	–222,4	114 582	3 171,8
Foobmaker	34,1	69,8	29 000	1 071,7
Host Marriott Services	20,8	45,5	24 000	1 284,6
McDonald's	1 642,5	4,4	237 000	11 408,8
Outback Steakhouse	61,5	–14,2	19 000	1 151,6
Performance Food Group	13,2	20,6	3 000	1 230,1
ProSource	–13,7	0,0	3 400	3 901,2
Shoney's	–35,7	–173,6	33 000	1 202,8
Viad	89,3	214,8	23 613	2 417,5
Wendy's International	130,5	–16,3	27 500	2 037,3

Источник данных: `http:/www.pathfinder.com/fortune/fortune500/ind147.html`; осень 1998 г.



Рассмотрим еще несколько примеров наборов многомерных данных.

1. Темп роста (выделенная переменная) и набор характеристик стратегии (остальные переменные), таких как тип оборудования, объем инвестиций, стиль руководства для каждой из нескольких новых предпринимательских фирм. Анализ мог бы показать, какое сочетание приводит к успеху, а какое — нет.
2. Заработная плата (выделенная переменная) а также пол (регистрируется как "мужской/женский", или 1/0), стаж работы, категория работы и производительность для каждого служащего. Такие данные могут рассматриваться в судебном процессе о дискриминации (с точки зрения более низкой средней оплаты труда) женщин. Ключевой вопрос, на который может ответить многомерный анализ, заключается в следующем. Можно ли объяснить расхождение в размере заработной платы факторами, отличными от пола служащего? Статистические методы могут исключить влияние этих остальных факторов и таким образом измерить среднее различие заработной платы между мужчинами и женщинами, которые одинаковы в других отношениях.
3. Для каждого из домов в районе цена этого дома (выделенная переменная) и ряд переменных, от которых зависит стоимость недвижимости, а именно количество домов такого типа, площадь дома, количество комнат, наличие или отсутствие бассейна, возраст дома. Анализ показал бы, как оценивается недвижимость в этом районе. Такой результат можно было бы использовать для определения реальной рыночной стоимости дома в этом районе или при строительстве, чтобы определить, какая комбинация характеристик нового дома повышает его цену.

2.2. Количественные данные: числа

Числа, имеющие содержательную интерпретацию, — это числа, которые непосредственно представляют измеренный или наблюдаемый *объем* некоторого признака или количество элементарных единиц. К числам, имеющим содержательную интерпретацию, можно отнести, например, количество долларов, частоты, размеры, количество служащих или число милей на галлон. К ним *не относятся* те числа, которые используют для кодирования или нумерации чего-либо, как, например, номер на футбольной спортивной форме или кодирование сделок вида 1 = покупка акции, 2 = продажа акции, 3 = покупка обязательств, 4 = продажа обязательств. Если данные представляют собой числа, имеющие содержательную интерпретацию, то мы имеем дело с **количественными** данными (т.е. они представляют количество чего-либо). С количественными данными можно выполнять все обычные операции над числами, такие как вычисление среднего (см. главу 4) и оценку изменчивости (см. главу 5). С такими данными можно проводить непосредственные вычисления. В зависимости от того, какие значения может потенциально принимать переменная, выделяют два типа количественных данных: *дискретные* и *непрерывные*.

Дискретные количественные данные

Дискретная переменная — это такая переменная, которая может принимать значения только из некоторого списка определенных чисел[2]. Например, число детей в семье является дискретной переменной. Поскольку возможные значения переменной можно перечислить, то с наборами дискретных данных работать относительно легко. Рассмотрим несколько примеров дискретных переменных.

1. Сколько раз за последние 24 часа на предприятии выключали компьютер.
2. Количество действительно заключенных контрактов из 18 подготовленных вами предложений.
3. Число иностранных танкеров, пришвартовавшихся сегодня в определенном порту.
4. Пол служащего, записанный с помощью числа 0 или 1.

Непрерывные количественные данные

Непрерывной будем считать любую числовую переменную, которая не является *дискретной*[3]. Слово "непрерывная" используют, поскольку возможные значения переменной образуют "континуум", как, например, множество всех положительных чисел, множество всех чисел или все значения между 0 и 100%. Например, фактический вес леденца на палочке, записанный как "нетто вес 1,7

[2] Обратите внимание на разницу между *дискретной (discrete)* переменной (рассматриваемой здесь) и *умеренной (discreet)* переменной, действие которой является достаточно мягким.

[3] Хотя такого определения достаточно для многих приложений в бизнесе, математическая теория более сложна и требует более тщательного определения, включающего интегральное исчисление (которое в этой книге не рассматривается). Мы также воздержимся от обсуждения смешанных переменных, которые не являются ни дискретными, ни непрерывными.

унции", представляет собой непрерывную случайную переменную, поскольку фактический вес может быть равен 1,70235 или 1,69481 унции, а не точно 1,7 унции. Если вы все еще не обладаете статистическим мышлением, можете считать, что фактический вес точно равен 1,7 унции; в действительности в любых реальных измерениях всегда есть небольшие (а иногда большие) отклонения от ожидаемых значений.

Рассмотрим несколько примеров непрерывных переменных.

1. Цена унции золота в долларах в настоящий момент. Казалось бы, что можно рассматривать эту величину как дискретную (и технически это верно, поскольку число вида 390,79 доллара принадлежит списку дискретных значений, записанных с точностью до центов: 0,00 , 0,01, 0,02,...). Однако лучше рассматривать такие числа как непрерывные данные, поскольку размер дискретного изменения мал и не важен для анализа. Если бы золото продавалось по цене несколько центов за унцию, то его стоимость, возможно, нужно было бы рассматривать как дискретные данные. Однако более вероятно, что в этом случае цена была бы указана с учетом тысячных долей цента, что, по существу, тоже выступало бы как непрерывное количество.
2. Инвестиционные показатели и показатели бухгалтерского учета: доход на одну акцию, ставка дохода на инвестиции, показатель перекрытия.
3. Количество энергии, необходимое для работы одного станка.

Остерегайтесь чисел, не имеющих содержательную интерпретацию

Прежде чем вы приступите к анализу количественных данных, следует сделать одно важное предупреждение. Убедитесь, что числа, которые вы анализируете, действительно имеют содержательную интерпретацию. К сожалению, числа можно использовать для записи чего угодно. Если вы имеете дело с произвольным кодированием, то результат анализа таких данных не будет иметь смысла.

Пример. *Алфавитный порядок штатов*

Предположим, что штаты США расположены в алфавитном порядке и закодированы как 1, 2, 3 . . . :

1	Алабама
2	Аляска
3	Аризона
4	Арканзас
.	.

Теперь представьте себе, что вас интересует вопрос о среднем штате из числа тех, в которых проживают служащие вашей фирмы. Конечно, формально ответ на такой вопрос вычислить можно. Однако результат будет лишен смысла, поскольку числа, которыми обозначены штаты, не имеют содержательного смысла (хотя эти числа можно использовать для других целей). Вряд ли среднее значение 28,35 или что-то между Невадой и Нью-Гемпширом может представлять интерес для вас. Мораль такова: прежде чем работать с числами, убедитесь, что они имеют содержательный смысл.

2.3. Качественные данные: категории

Если набор данных показывает, какой из нескольких нечисловых категорий принадлежит каждый из объектов, то данные являются **качественными** (поскольку они регистрируют определенное качество, которым обладает объект). Будьте внимательны и осторожны, чтобы избежать искушения приписать числовые значения категориям (классам) и далее проводить с ними вычисления. Если имеется несколько классов, то можно оперировать процентами (частотами) событий в каждом классе (создав таким образом нечто числовое из представленных категориями данных). Если есть в точности две категории, их можно обозначить цифрами 1 и 0, приписать эти значения соответственно каждому из объектов и затем (в достаточно многих случаях) обрабатывать полученные данные как количественные. Давайте сначала рассмотрим общий случай, когда речь идет о трех или более категориях.

Существуют два типа качественных данных: *порядковые* (ординальные, для которых существует имеющий содержательный смысл порядок, но нет содержательного числового обозначения) и *номинальные* (для которых нет содержательно интерпретируемого порядка).

Порядковые качественные данные

Набор данных является **ординальным**, если существует имеющий содержательный смысл порядок: можно вести речь о первом (например, "лучшем"), втором, третьем и т.д. Можно ранжировать данные в соответствии с этим порядком и использовать это ранжирование при выполнении анализа, особенно если оно имеет отношение к изучаемому вопросу. В главе 4 будет рассмотрена *медиана* (среднее значение упорядоченного ряда), как пример статистического показателя.

Рассмотрим некоторые примеры порядковых данных.

1. Должность, записанная для каждого из группы руководителей: президент, вице-президент, начальник отдела, заместитель начальника отдела. Хотя классификатор не содержит чисел и не совсем ясно, каким образом их можно здесь использовать, объекты можно естественным образом упорядочить.
2. Характеристики, такие как AA+, AA, AA–, A+, A, A–, B+, B и B–, зафиксированные для набора долговых обязательств Это чисто порядковые категориальные данные, поскольку упорядоченность имеет смысл с точки зрения риска вкладов и используется в анализе инвестиций.
3. Ответы на вопрос анкеты: "Пожалуйста, выскажите свое мнение относительно вашей работы в фирме, используя шкалу от 1 до 5, где 1 означает "с трудом дожидаюсь окончания рабочего дня", а 5 — "все мои мысли заняты работой". Несмотря на то что ответы выражены числами, мы имеем дело с порядковыми данными, поскольку предложенная шкала оценок носит субъективный характер. Непонятно, можно ли считать, что разница между оценками 5 и 4 такая же, как и между оценками 2 и 1. Кроме того,

можно ли считать, что оценка 2 в два раза лучше оценки 1. Однако упорядочение и ранжирование здесь явно имеют место [4].

Номинальные качественные данные

Номинальные качественные данные определяются в терминах категорий, которые нельзя содержательно упорядочить. Для таких категорий нет чисел, с которыми можно производить вычисления, и нет основы для ранжирования. Все, что можно сделать, — это подсчитать процент (или количество) попадающих в каждую из категорий наблюдений и использовать в качестве обобщающего показателя *моду* (наиболее часто встречающаяся категория; формально этот показатель будет определен в главе 4).

Рассмотрим несколько примеров номинальных данных.

1. Штаты, в которых проживают служащие вашей фирмы. Мы уже отмечали раньше, что это лишь категории. Чтобы ранжировать штаты, необходимо обязательно ввести какую-либо другую переменную (например, население штата или размер дохода на душу населения), которую лучше использовать непосредственно.
2. Главный продукт каждого из нескольких производственных предприятий диверсифицированного бизнеса, как, например, пластмасса, электроника, древесина. Эти категории действительно не упорядочены. Чтобы их упорядочить, необходимо рассмотреть дополнительный фактор (как, например, потенциал роста данной фирмы в отрасли), не являющийся внутренним свойством этих категорий.
3. Названия всех фирм, указанных на первой странице сегодняшнего выпуска *The Wall Street Journal*.

2.4. Временны*е* ряды и данные об одном временно*м* срезе

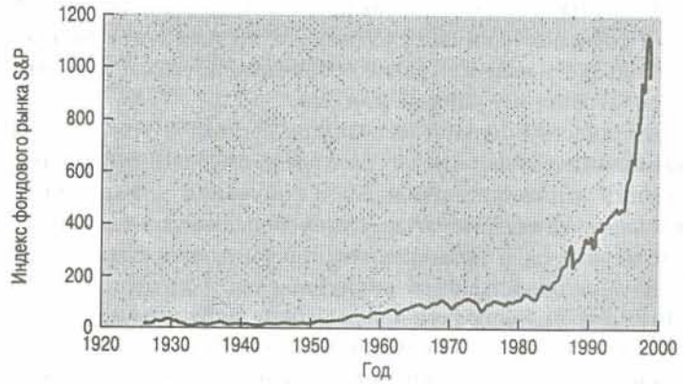
Если порядок записи значений данных имеет содержательный смысл, как, например, ежедневные цены на фондовом рынке, то мы имеем дело с **временны*м* рядом.** Если последовательность, в которой записаны данные, не важна, как, скажем, доходы восьми аэрокосмических фирм в первом квартале 1996 года, то мы имеем данные **об одном временном срезе.** Слова *об одном временном срезе* в данном случае означают лишь то, что нет никакого упорядочения во времени, а есть лишь информация о некоторых объектах в определенный момент времени (своего рода "моментальный снимок").

Анализ временны*х* рядов в целом сложнее, чем анализ данных об одном временно*м* срезе, поскольку требует тщательного учета порядка наблюдений. Поэтому в следующих главах мы начнем с данных об одном временно*м* срезе. Временны*е* ряды будут рассмотрены позже, в главе 14.

[4] Внимательный читатель мог бы спросить, действительно ли оценка 4, данная одним человеком, обязательно выше, чем оценка 3, данная другим. Другими словами, из-за отсутствия стандартизации человеческого восприятия мы можем не получить даже действительно порядковую шкалу. Эта одна их многих сложностей статистического анализа, которую мы не можем учесть в полной мере.

Пример. *Фондовый рынок*

Ниже приведен график индекса фондового рынка S&P, цены на момент окончания торгов на бирже в конце каждого месяца, начиная с декабря 1925 года. Временной ряд показывает изменение стоимости портфеля ценных бумаг с течением времени. Обратите внимание, как впечатляюще поднялась стоимость акций на фондовом рынке в последние годы, хотя этот процесс был не совсем гладким; направленные вниз зубцы графика (например, крах фондового рынка в 1987 году) показывают риск держателя портфеля акций, которые обычно (но не всегда) растут в цене.



Рассмотрим еще несколько примеров временных рядов.

1. Цена на пшеницу за последние 50 лет с учетом инфляции. Считая, что изменения в будущем будут носить тот же характер, что и изменения в прошлом, можно использовать эти временны*е* тренды для долгосрочного планирования.
2. Объемы месячных продаж за последние 20 лет. Этот набор данных имеет структуру, показывающую рост продаж с течением времени, а также отчетливую сезонную особенность с пиками в декабрьские праздники.
3. Результаты ежеминутных измерений толщины бумаги на выходе из бумагоделательной машины. Такие данные важны для контроля качества. Временна*я* последовательность важна, поскольку небольшие изменения толщины бумаги могут либо последовательно "дрейфовать" в сторону недопустимого уровня, либо "колебаться", становясь шире или у*же* в четко допустимых пределах.

Теперь рассмотрим несколько примеров наборов данных об одном временном срезе.

1. Измеренная для 30 человек продолжительность сна в последнюю ночь, которая используется для оценки эффективности нового лекарства, продаваемого без рецепта.
2. Сегодняшняя балансовая стоимость случайной выборки банковских сберегательных сертификатов.

3. Количество телефонных звонков, обработанных вчера каждым из работающих с заказами служащих фирмы.

2.5. Источники данных, включая Internet

Откуда берут данные? Существует много источников, выбор которых осуществляют исходя из их стоимости, доступности и потребностей экономической деятельности. Если вы самостоятельно разрабатываете план сбора данных (даже если собственно собирают данные другие), то вы получите **первичные данные**. Если же вы используете данные, ранее собранные другими людьми и для других целей, то вы используете **вторичные данные**.

Главное преимущество первичных данных состоит в том, что в этом случае у вас больше возможностей собрать действительно необходимую вам информацию, поскольку *вы* сами управляете процессом получения данных путем планирования вопросов или измерений, а также путем определения выборки элементарных единиц для измерения. К сожалению, часто получение первичных данных слишком дорого и занимает много времени. С другой стороны, вторичные данные дешевле (или вообще бесплатные), и можно найти именно то (или почти то), что нужно. Это предполагает следующую стратегию получения данных: поиск вторичных данных, которые быстро удовлетворяют ваши потребности за приемлемую цену. Если это невозможно, оцените стоимость сбора первичных данных и решайте, какой источник (первичный или вторичный) использовать, исходя из соотношения расходов и преимуществ каждого из подходов.

Рассмотрим несколько примеров источников первичных данных.

1. Информация о производительности вашего оборудования, включая объем и качество (например, уровень брака) ежедневно выпускаемой продукции. Такие данные может автоматически собирать информационная система вашей компании.
2. Данные опроса, проведенного служащими маркетинговой фирмы, нанятыми вами с целью изучения влияния возможной рекламной кампании на поведение потребителей.
3. Собранные в ходе политической кампании данные о проблемах, которыми обеспокоены избиратели, собирающиеся голосовать на предстоящих выборах.

А теперь рассмотрим примеры источников вторичных данных.

1. Собранные и сведенные в таблицу правительством США экономические и демографические данные, которые доступны бесплатно в библиотеке или через Internet.
2. Данные из специализированных журналов (например, реклама, объемы производства, финансы и т.п.), которые помогают фирмам, работающим в этом секторе рынка, оценить ситуации на рынке и успех отдельных продуктов.
3. Данные, собранные компаниями, специализирующимися на сборе данных и продающими их другим компаниям. Например, Nielsen Media Research продает телевизионные рейтинги (исходя из наблюдений за тем, какие телешоу смотрела выборка людей) телевизионным кабельным сетям, независимым станциям, рекламодателям, рекламным агентствам и др. Большинство публикаций, которые можно найти в библиотеке, ссылаются именно на специализированные данные, полученные такими компаниями.

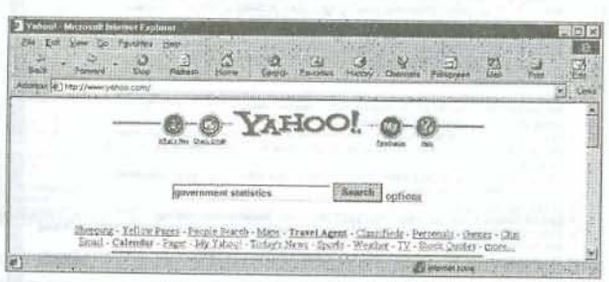


Для поиска данных в Internet большинство людей используют поисковые системы (например, Yahoo!, расположенную по адресу `http://www.yahoo.com`), задавая определенные ключевые слова (например, "правительственная статистика", "финансовые данные"). Поисковая система выдает перечень ссылок на расположенные в разных частях мира электронные страницы (сайты), содержание которых связано с тем, что вы ищете. Чтобы попасть на сайт, содержимое которого, возможно, представляет для вас интерес, щелкните на соответствующей ссылке (обратите внимание, что адреса в Internet с окончанием `.com` представляют коммерческие сайты, с окончанием `.gov` — государственные, а с окончанием `.edu` относятся к сфере образования), а затем щелкните на кнопке Back, чтобы вернуться к списку и выбрать другой сайт. Сеть Internet сильно изменилась за последние несколько лет, и, похоже, что в ней накапливается все больше полезной и доступной информации. В частности, почти все крупные фирмы сейчас имеют свои собственные сайты. К сожалению, достаточно часто таким образом не удается найти ту информацию, которая действительно нужна.

Пример. *Поиск в Internet данных государственной статистики о потребительских ценах*

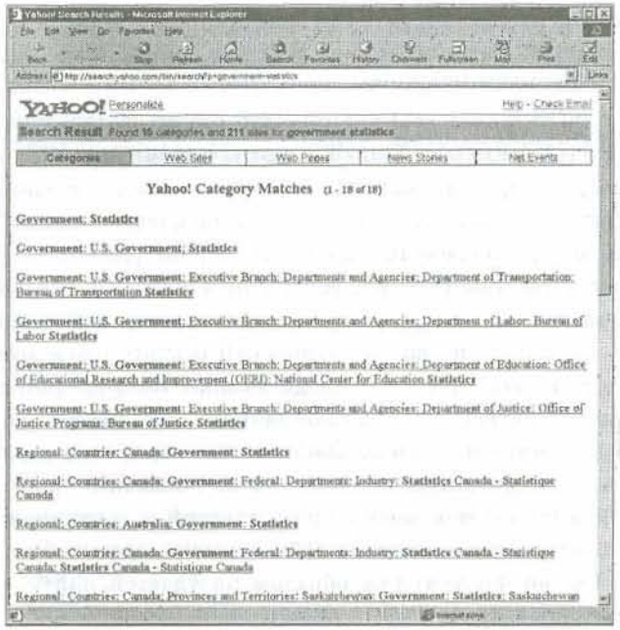
Internet стала большим источником различного рода информации. Давайте посмотрим, как можно выполнить поиск данных государственной статистики и скопировать их в электронные таблицы Microsoft Excel для дальнейшего анализа.

Начните работу, указав программе просмотра (броузеру), например Microsoft Internet Explorer или Netscape Navigator, связаться с вашей любимой поисковой системой. Для этого введите, например, `www.yahoo.com` в поле адреса или щелкните на пиктограмме Search (Поиск). Затем введите в поле поиска ключевые слова (в нашем случае — government statistics). Экран вашего компьютера будет выглядеть следующим образом[5].

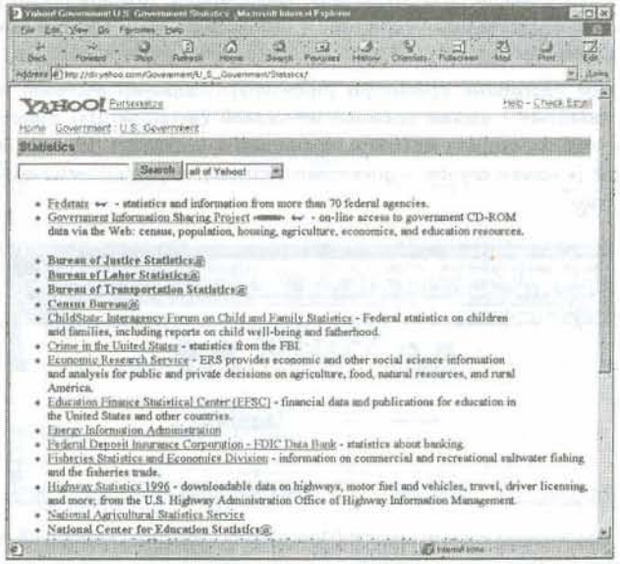


Чтобы начать поиск, щелкните на кнопке Search (Поиск) (или нажмите клавишу <Enter>). После этого вы получите длинный перечень (возможно) необходимых вам электронных страниц. В нашем случае информация о 18 соответствующих категориях и 211 отдельных страницах будет размещена на экране так, как показано на рисунке.

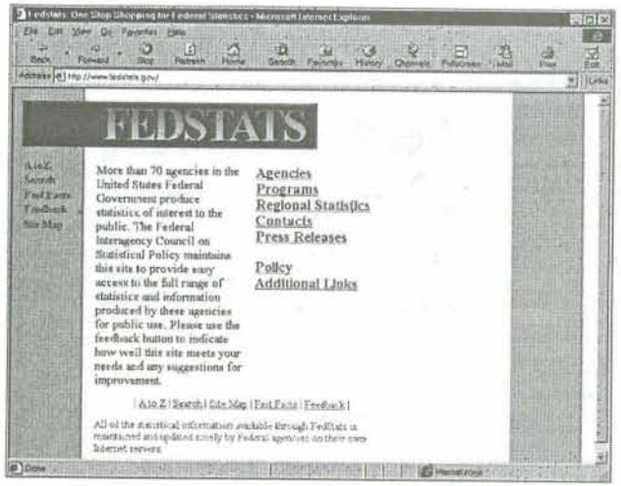
[5] Изображение на вашем экране может отличаться от приведенного здесь по ряду причин, в том числе и из-за изменения самой Internet.



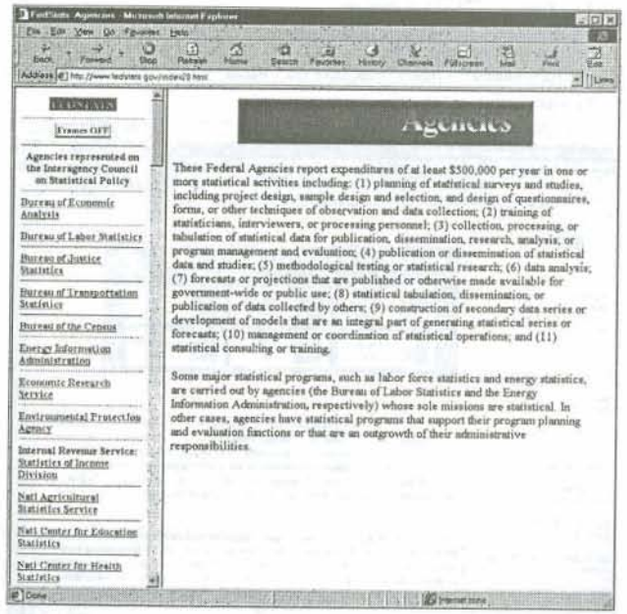
Выбрав "Government: U.S. Government: Statistics" в верхней части списка (просто щелкнув мышью на подчеркнутом тексте), вы получите подробную информацию, в частности сведения о государственной статистике США.



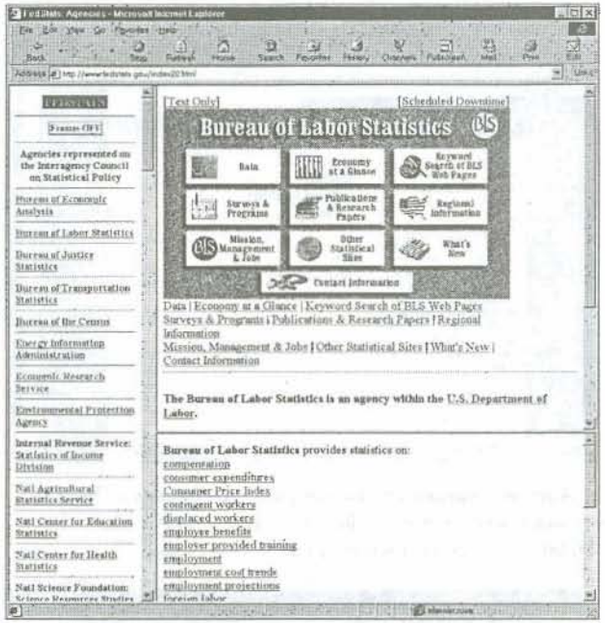
Выбрав "Fedstats", первую ссылку в этом списке, вы получите список правительственных учреждений США, которые "готовят статистические данные для общественности".



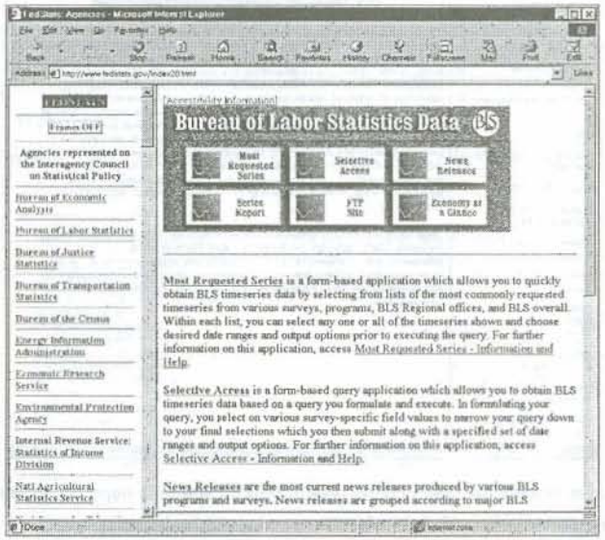
Далее, выбрав ссылку "Agencies" вверху справа, как раз под названием "Fedstats", вы получите длинный список учреждений в рамке слева (например, Bureau of Economic Analysis — Бюро экономического анализа, Bureau of Labor Statistics — Бюро статистики труда...).



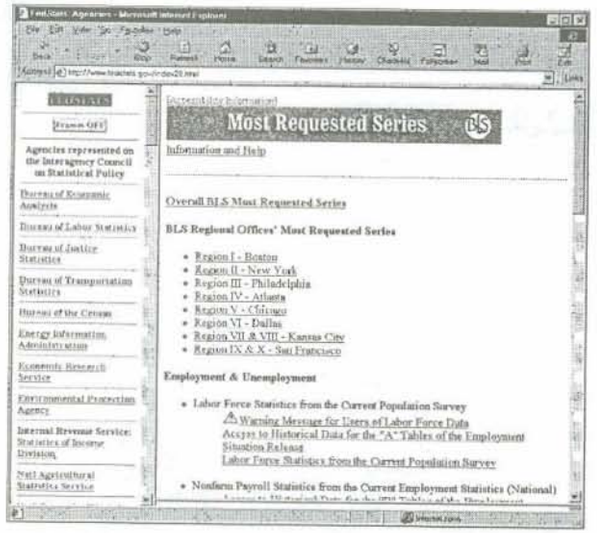
Выбрав слева на экране ссылку на Бюро статистики труда в середине перечня учреждений, вы получите информацию, которую предлагает это учреждение.



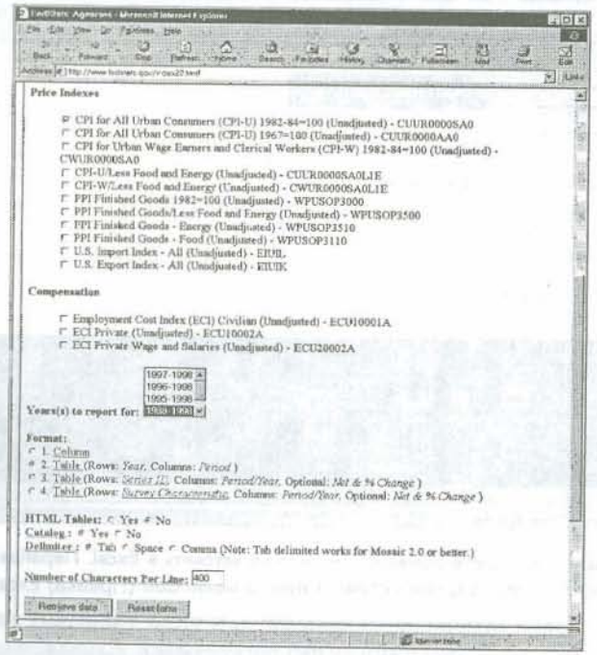
Щелкнув на пиктограмме Data вверху слева, как раз под названием "Bureau of Labor Statistics", вы получите информацию о доступных данных, как показано ниже.



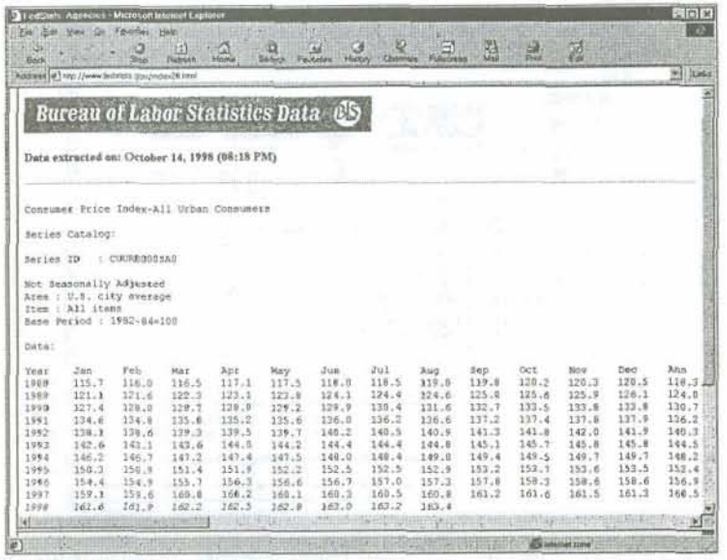
Щелкнув на ссылке "Most Requested Series" (Наиболее запрашиваемые данные) в середине этой страницы, вы получите следующие варианты.



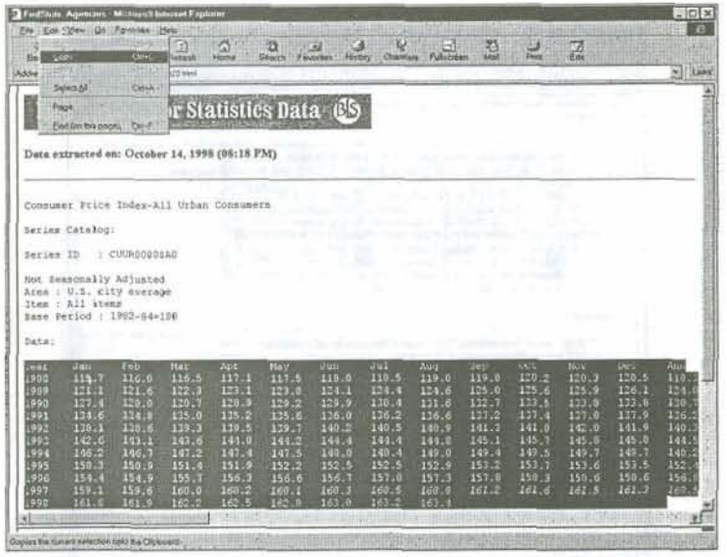
Выбрав "Overall BLS Most Requested Series" (Полный список наиболее запрашиваемых данных Бюро статистики труда), прокрутите его вниз, запросите данные относительно первого индекса цен (Индекс потребительских цен — "Consumer Price Index" (CPI)), выберите период времени 1988–1998, отметьте "No" рядом с "HTML Tables" (HTML Таблицы) (чтобы позже можно было легко скопировать данные в Excel), отметьте в качестве ограничителя "Tab" (по той же причине), и вы получите следующее.



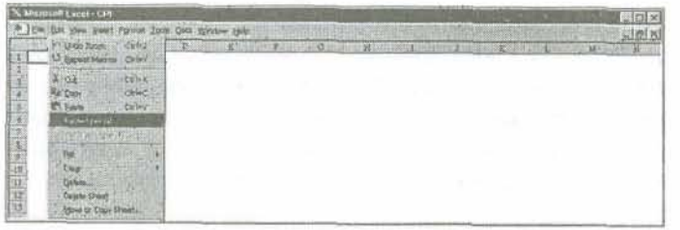
Щелкнув на кнопке Retrieve data (Запрос данных), вы получите следующий набор данных, состоящий из значений индексов потребительских цен по месяцам (и по годам — в крайней правой колонке).



Перемещая мышку с расположенного в левом верхнем углу слова "Year" вниз в правый угол, выделите данные и скопируйте их с помощью команды Copy (Копировать) пункта Edit (Правка) главного меню следующим образом.



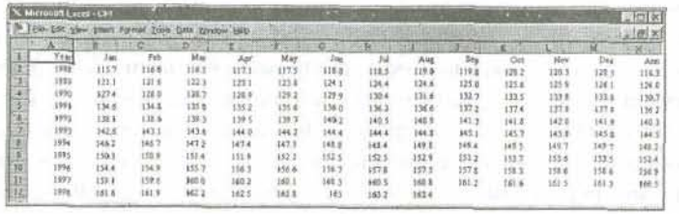
Теперь данные находятся в буфере обмена, и их можно вставить в Excel. Перейдя в Excel, используйте команду Paste Special (Специальная вставка) пункта меню Edit (Правка) Excel, и вы получите следующее.



Затем выберите в появившемся диалоговом окне пункт Text (Текст), чтобы каждое число набора данных попало в отдельную ячейку.



В результате ваш набор данных (после дополнительного форматирования таким образом, чтобы значения были представлены одной цифрой после десятичной точки и выровнены вправо) появится в Excel и будет готов для анализа.*



Поздравляю! Данные успешно найдены и помещены в электронную таблицу. Конечно, все эти шаги необходимо *повторить несколько раз для разных ссылок в надежде найти интересующие вас данные.* При этом приходится изменять ключевые слова и поисковую систему, возвращаться назад с помощью кнопки Back (Назад), проверяя различные ссылки.

*Примечание. Если Excel поместит все ваши данные в одну колонку (вместо того, чтобы поместить каждое число в отдельную ячейку), выделите колонку, выберите команду Data⇨TextToColumns (Данные⇨Текст по столбцам) из меню Excel, затем — опцию Delimited (С разделителями) в окне мастера преобразования текстов и попробуйте правильно упорядочить данные, выбирая в качестве разделителя пробел и/или символ табуляции.

2.6. Дополнительный материал

Резюме

Набор данных содержит одно или несколько значений для каждого из отдельных объектов, называемых **элементарными единицами**. В качестве таких объектов могут выступать люди, домохозяйства, города, телевизионные приемники или что угодно, что представляет интерес для изучения. Для каждого из объектов регист-

рируют один и тот же признак (или признаки). Признак, который регистрируют для каждого из объектов (например, стоимость), называется **переменной**.

Существуют три основных способа классификации наборов данных: по количеству переменных (одномерный, двумерный и многомерный); по типу представленной каждой из переменных информации (числа или категории) и в зависимости от того, является ли набор данных временным рядом, или это данные об одном временном срезе.

Одномерные наборы данных (одна переменная) содержат информацию только об одном признаке, зарегистрированную для каждого объекта. Одномерный набор данных позволяет определить типичное значение и характеристику изменчивости данных, а также выделить специфические особенности или проблемы в данных.

Двумерные наборы данных (две переменные) содержат два признака, значения которых регистрируются для каждого объекта. Двумерные данные в дополнение к информации о каждой переменной как наборе одномерных данных позволяют изучить связь между двумя переменными и предсказать значение одной переменной на основе значения другой.

Многомерные наборы данных (много переменных) содержат три или больше признаков, значения которых регистрируются для каждого объекта. Многомерные данные в дополнение к информации о каждой переменной как наборе одномерных данных дают возможность изучить связь между переменными и предсказать значение одной переменной на основе значения других.

Значения переменных, которые регистрируются как числа, имеющие содержательный смысл, называют **количественными** данными. **Дискретная** количественная переменная может принимать значения только из некоторого списка конкретных чисел (таких как, например, 0 или 1, или перечень чисел 0, 1, 2, 3, . . .). Любую количественную переменную, которая не является дискретной, будем называть **непрерывной**. Значения непрерывной переменной не ограничены простым перечнем возможных значений.

Если переменная содержит информацию о том, какой из нескольких нечисловых категорий принадлежит объект, то она называется **качественной** переменной. Если категории можно естественным образом и содержательно осмысленно упорядочить, то речь идет о **порядковой** качественной переменной. Если такой порядок отсутствует, то речь идет о **номинальной** качественной переменной. Несмотря на то что часто значения качественной переменной можно записать с помощью чисел, такая переменная все равно остается качественной, а не количественной, поскольку эти числа не имеют какой-либо интерпретации, содержательно присущей этой переменной.

К количественным данным можно применять те же операции, что и к обычным числам: подсчет частоты, ранжирование, арифметические действия. С порядковыми данными можно выполнять только подсчет частоты и ранжирование, с номинальными данными — только подсчет частоты.

Если последовательность записи данных имеет содержательный смысл, то соответствующий набор данных представляет собой **временной ряд**. Если последовательность записи данных не важна, то соответствующий набор содержит данные **об одном временном срезе**. Анализ временных рядов сложнее анализа данных об одном временном срезе.

Если вы самостоятельно планируете сбор данных (даже если собственно сбор данных делают другие), то получаете **первичные данные**. Если вы используете данные, предварительно собранные другими людьми и для других целей, то вы имеете дело с **вторичными данными**. Получение первичных данных часто обходится дорого и занимает много времени, но вы получаете то, что вам необходимо. Вторичные данные можно получить дешевле (или даже бесплатно), но вы можете найти, а можете и не найти то, что вам необходимо.

Основные термины

- Набор данных (data set), 42
- Элементарные единицы (elementary units), 42
- Переменная (variable), 42
- Одномерный (univariate), 43
- Двумерный (bivariate), 44
- Многомерный (multivariate), 45
- Количественная (quantitative), 47
- Дискретная (discrete), 47
- Непрерывная (continuos), 47
- Качественная (qualitative), 49
- Порядковая или ординальная (ordinal), 49
- Номинальная (nominal), 50
- Временные ряды (time series), 50
- Об одном временном срезе (cross-sectional), 50
- Первичные данные (primary data), 52
- Вторичные данные (secondary data), 52

Контрольные вопросы

1. Что такое набор данных?
2. Что такое переменная?
3. Что такое элементарная единица?
4. Какими тремя основными способами можно классифицировать наборы данных? (*Подсказка*. Не отвечайте просто "одномерный, двумерный, многомерный". Дайте развернутый ответ.)
5. На какие основные вопросы можно ответить, проанализировав:
   а) Одномерные данные?
   б) Двумерные данные?
   в) Многомерные данные?
6. Почему двумерные данные представляют собой больше, чем просто два отдельных одномерных набора данных?

7. Что можно делать с многомерным набором данных?
8. В чем разница между качественными и количественными данными?
9. В чем разница между дискретной и непрерывной количественными переменными?
10. Что представляют собой качественные данные?
11. В чем разница между порядковыми и номинальными качественными данными?
12. Чем отличаются временные ряды от данных об одном временном срезе?
13. Что легче анализировать: временные ряды или данные об одном временном срезе?
14. Определите разницу между первичными и вторичными данными.

Задачи

1. Назовите два различных набора двумерных данных, которые прямо или косвенно связаны с вашими обязанностями. В каждом случае определите характер зависимости между двумя факторами и установите, можно ли и полезно ли иметь возможность предсказывать один фактор на основании другого.
2. Выполните предыдущую задачу, но для многомерных данных.
3. Выберите некоторую фирму и назовите две количественные переменные, представляющие интерес для нее. Укажите для каждой переменной, является она дискретной или непрерывной.
4. Выберите некоторую фирму и назовите две качественные переменные, представляющие для нее интерес. Для каждой переменной укажите, является она номинальной или порядковой.
5. Укажите три временных ряда, представляющих для вас интерес. Для каждого ряда определите:

   а) Есть ли временной тренд?

   б) Есть ли сезонные влияния?
6. Выберите некоторую фирму и опишите (в общих терминах) базу данных этой фирмы. Внутри этой базы данных определите три набора данных различных видов. Для каждого из этих трех наборов укажите, что является элементарной единицей, и определите, что можно узнать в результате соответствующего анализа.
7. Ваша фирма решила подать в суд на ненадежного поставщика. Какой вид анализа можно выполнить для оценки потерь из-за упущенных возможностей, исходя из результатов конкурентов, экономической ситуации и времени года?
8. Определите вид (первичные или вторичные) следующих данных.

   а) Данные правительства США о текущей экономической ситуации в каждом из штатов, используемые фирмой, планирующей расширение.

   б) Данные о себестоимости продукции одного из предприятий вашей фирмы, собранные в ходе кампании по снижению затрат.

   в) Данные отчета по отрасли, приобретенные вашей фирмой с целью оценки своего места среди конкурирующих фирм.
9. В табл. 2.6.1 содержится несколько объектов из базы данных сотрудников. Информация дана для 5 человек по состоянию на 3 июля 1999 года.

   а) Что является элементарной единицей в этом наборе данных?

   б) Определите вид данных: одномерные, двумерные, многомерные?

   в) Какие из этих четырех переменных являются качественными, а какие — количественными?

   г) Какие из переменных (если такие есть) являются порядковыми качественными переменными? Поясните свой ответ.

   д) Это временной ряд или это данные об одном временном срезе?
10. Рассмотрим набор данных из табл. 2.6.2, содержащий информацию о пяти цехах (обозначаются кодами).

    а) Что является элементарной единицей для этого набора данных?

    б) Определите вид данных: одномерные, двумерные, многомерные?

    в) Укажите качественные переменные (если они есть).

    г) Есть ли в этих данных порядковая переменная? Если да, то укажите ее.

    д) Это временной ряд или данные об одном временном срезе?
11. В табл. 2.6.3 содержатся данные об объемах продаж и размерах дохода (в тысячах долларов); информация за первые 6 месяцев 1998 года.

Таблица 2.6.1. Данные о пяти служащих

Пол	Зарплата, дол.	Образование	Стаж, лет
М	42 300	Высшая школа	9
Ж	31 800	Колледж	4
М	29 500	Степень магистра	2
Ж	58 100	Степень магистра	15
Ж	36 000	Колледж	7

Таблица 2.6.2. Информация о некоторых видах продукции пяти цехов

Код	Деталь	Качество	Количество служащих
A-235-86	Тормоза	Хорошее	53
W-186-74	Топливопровод	Отличное	37
X-937-85	Радио	Довольно хорошее	26
C-447-91	Шасси	Превосходное	85
F-258-89	Провод	Хорошее	16

а) Что является элементарной единицей для этого набора данных?

б) Определите вид данных: одномерные, двумерные, многомерные?

в) Какая из двух переменных является количественной, а какая переменная является качественной?

г) Это временной ряд или данные об одном временном срезе?

12. Табл. 2.6.4 представляет собой часть базы данных о потребителях одного из сотрудников отдела продаж.

а) Что является элементарной единицей в этом наборе данных?

б) Определите вид данных: одномерные, двумерные, многомерные?

в) Определите, какие из этих переменных количественные, а какие качественные?

г) Какие из этих переменных номинальные? Какие порядковые?

д) Это временной ряд или данные об одном временном срезе?

13. Чтобы спланировать объем затрат на рекламу в различных средствах массовой информации (телевидение, радио, газеты и др.), вы изучаете набор данных, содержащий прошлогодние расходы каждого из ваших конкурентов на телерекламу, радиорекламу и на рекламу в газетах. Дайте полное описание типа для такого набора данных.

14. Объемы квартальных продаж вашей фирмы за последние пять лет могут быть полезны для стратегического планирования.

а) Это временной ряд или данные об одном временном срезе?

б) Определите вид данных: одномерные, двумерные, многомерные?

15. Рассмотрим информацию о предлагаемой и запрашиваемой цене для 18 различных обязательств Казначейства США на момент закрытия торгов в конкретный день.

а) Определите вид данных: одномерные, двумерные, многомерные?

б) Это временной ряд или данные об одном временном срезе?

16. Рассмотрим данные продаж 35 компаний.

а) Определите вид данных: одномерные, двумерные, многомерные?

б) Это качественная или количественная переменная?

в) Это порядковая, номинальная или какая-либо другая переменная?

17. Инспектор по контролю качества оценил каждую из произведенных сегодня партий продукции по шкале от A до E, где A — высший сорт, E — низший.

а) Какая это переменная: количественная или качественная?

б) Это порядковая, номинальная или какая-либо другая переменная?

18. Одна из колонок электронной таблицы содержит названия компаний, поставляющих вам каждое из комплектующих.

а) Какая это переменная: количественная или качественная?

б) Это порядковая, номинальная или какая-либо другая переменная?

19. В табл. 2.6.5 содержатся рейтинги качества лезвий для бритв.

а) Что является элементарной единицей в этом наборе данных?

б) Это одномерные, двумерные или многомерные данные?

в) Это временной ряд или данные об одном временном срезе?

Таблица 2.6.3. Продажи и доход с января по июнь 1998 года

Продажи	Доход (убытки)
350	30
270	23
140	(2)
280	14
410	53
390	47

Таблица 2.6.4. Некоторые потребители и покупки

Уровень интереса к новым изделиям	Общий объем закупок в прошлом году, дол.	Географический регион
Слабый	88 906	Запад
Умеренный	396 808	Юг
Очень сильный	438 442	Юг
Слабый	2 486	Средний запад
Слабый	37 375	Запад
Очень сильный	2 314	Северо-восток
Умеренный	1 244 096	Средний запад
Слабый	857 248	Юг
Сильный	119 650	Северо-восток
Умеренный	711 514	Запад
Слабый	22 616	Запад

Таблица 2.6.5 Рейтинги качества лезвий для бритв

Изделие	Тип	Цена, дол.	Удобство в обращении
Gillette Sensor Excel	Бритвенный блок	4,50	Отличное
Schick Tracer	Бритвенный блок	3,34	Очень хорошее
Bic Twin Select for Sensible Skin	Одноразовые	3,68	Хорошее
Gillette Sensor for Women	Бритвенный блок	3,93	Отличное
Schick Silk Effects for Women	Бритвенный блок	4,64	Очень хорошее
Bic Twin Pastel	Одноразовые	2,03	Хорошее

Данные взяты из журнала *Consumer Reports*, October, 1995, p. 649.

г) Переменная "тип" — количественная, порядковая или номинальная?

д) Переменная "цена" — количественная, порядковая или номинальная?

е) Переменная "удобство" — количественная, порядковая или номинальная?

20. Предположим, что набор данных включает переменную "тип ценных бумаг", закодированную таким образом: 1 — обыкновенная акция; 2 — привилегированная акция; 3 — обязательство; 4 — фьючерсный контракт; 5 — опцион. Это количественная или качественная переменная?

21. Технологичность сборки изделий оценивается по следующей шкале: 1 — высокотехнологичная сборка; 2 — технологичная; 3 — удовлетворительная; 4 — сложная; 5 — очень сложная. Это количественная, порядковая или номинальная переменная?

22. Предположим, что набор данных включает переменную "экономическая организация", закодированную следующим образом: 1 = один собственник, 2 = товарищество, 3 = S-корпорация, 4 = C-корпорация. Это количественная или качественная переменная?

23. В табл. 2.6.6 содержатся данные о бытовых пылесосах.

а) Что является элементарной единицей в этом наборе данных?

б) Это одномерные, двумерные или многомерные данные?

в) Какие из переменных являются качественными, а какие количественными?

г) Для каждой качественной переменной в этом наборе данных определите ее тип: порядковая или номинальная?

д) Это временной ряд или данные об одном временном срезе?

24. Компания Dow Jones рассчитывает ряд индексов фондового рынка, которые используются на Нью-йоркской фондовой бирже. Наиболее известен индекс Dow Jones Industrial Average (DJIA), который рассчитывают на основе данных об акциях 30 промышленных компаний. Другой индекс, называемый Dow Jones Transportation Average (DJTA), рассчитывают на основе данных об акциях 20 компаний, которые специализируются на оказании транспортных услуг (железнодорожный, автомобильный, морской транспорт, авиалинии и др.). В табл. 2.6.7 содержатся данные о 20 транспортных компаниях, которые фигурируют в DJTA.

Таблица 2.6.6 Сравнение пылесосов

Цена, дол.	Вес, фунты	Качество	Тип
170	17	Хорошее	Жесткий шланг
260	17	Относительно хорошее	Мягкий шланг; самоходный
100	21	Хорошее	Мягкий шланг
90	14	Хорошее	Жесткий шланг
340	13	Отличное	Мягкий шланг
120	24	Хорошее	Мягкий шланг; самоходный
130	17	Относительно хорошее	Мягкий шланг; самоходный

а) Что является элементарной единицей в этом наборе данных?

б) Это одномерные, двумерные или многомерные данные?

в) Какие из переменных являются качественными, а какие количественными?

г) Если в этом наборе есть качественные переменные, то какие они — порядковые или номинальные?

д) Это временной ряд или данные об одном временном срезе?

25. Продолжим работать с Dow Jones Transportation Average (DJTA), о котором речь шла в задаче 24. В табл. 2.6.8 содержатся данные о ежедневных наблюдениях значений DJTA, об изменении значения индекса между двумя последовательными наблюдениями и о процентном изменении DJTA между двумя последовательными наблюдениями.

Таблица 2.6.7. Цена акции и месячное процентное изменение для компаний, фигурирующих в DJTA

Компания	Цена на момент окончания торгов на бирже, 30 сентября 1998 года	Процентное изменение цены по сравнению с 31 августа 1998 года
Yellow Corp.	13,5	13,1
Southwest Airlines	20,0	12,3
CSX	42,1	11,4
UAL	64,8	7,5
Union Pacific	42,6	7,1
Ryder Systems	24,9	5,6
Burl. North. SF	32,0	3,2
Norfolk Southern	29,1	3,1
AMR	55,4	1,7
GATX Corp.	33,1	0,2
Delta Airlines	97,3	–4,7
CNF	29,1	–6,8
FDX Corp.	45,1	–9,9
Airborne Freight	17,3	–11,2
US Freightways	19,9	–11,4
Alaska Air	34,1	–12,5
US Airways	50,6	–13,1
Alex&Baldwin	19,9	–15,4
Xtra Corp.	46,6	–20,9
Roadway Express	11,0	–27,3

Данные взяты из The Wall Street Journal, September 1, 1998, p. C10A; October 1, 1998, p. C3.

а) Что является элементарной единицей в этом наборе данных?

б) Это одномерные, двумерные или многомерные данные?

в) Какие из переменных являются качественными, а какие количественными?

г) Если в этом наборе есть качественные переменные, то какие они — порядковые или номинальные?

д) Это временной ряд или данные об одном временном срезе?

Таблица 2.6.8. Ежедневные значения и изменения DJTA в сентябре 1998 года

Дата	DJTA	Изменение значения	Процентное изменение
31 августа	2752,51	–	–
1 сентября	2753,77	1,26	0,05
2 сентября	2702,55	–51,22	–1,86
3 сентября	2624,47	–78,08	–2,89
4 сентября	2616,75	–7,72	–0,29
8 сентября	2749,30	132,55	5,07
9 сентября	2691,20	–58,10	–2,11
10 сентября	2631,51	–59,69	–2,22
11 сентября	2679,17	47,66	1,81
14 сентября	2805,14	125,97	4,70
15 сентября	2863,70	58,56	2,09
16 сентября	2884,13	20,43	0,71
17 сентября	2813,99	–70,14	–2,43
18 сентября	2814,67	0,68	0,02
21 сентября	2802,64	–12,03	–0,43
22 сентября	2865,29	62,65	2,24
23 сентября	2904,10	38,81	1,35
24 сентября	2831,24	–72,86	–2,51
25 сентября	2795,83	–35,41	–1,25
28 сентября	2803,10	7,27	0,26
29 сентября	2746,13	–56,97	–2,03
30 сентября	2644,67	–101,46	–3,69

Данные взяты из The Wall Street Journal, разные номера за1998, с. A1

Упражнения с использованием базы данных

Обратитесь к базе данных о сотрудниках фирмы, содержащейся в приложении А.

1. Опишите и классифицируйте эту базу данных и ее части.

   а) Это одномерные, двумерные или многомерные данные?

   б) Что является элементарной единицей в этом наборе данных?

   в) Какие переменные качественные, а какие количественные?

   г) Переменная "уровень подготовки" является порядковой или номинальной? Почему?

   д) Можно ли выполнять арифметические действия с переменной "номер служащего"? О чем это свидетельствует с точки зрения того, является ли эта переменная действительно количественной?

   е) Это временной ряд или данные об одном временном срезе?

2. Для каждой переменной из этой базы данных определите, какие из указанных ниже операций можно применять к этой переменной.

   а) Арифметические (сложение, вычитание и т. д.).

   б) Подсчет количества служащих в каждой категории.

   в) Ранжирование по порядку.

   г) Вычисление процента служащих в каждой категории.

Проекты

1. Найдите в Internet, в статье из экономического журнала или в газете таблицу данных. Скопируйте эту статью и таблицу.

   а) Классифицируйте набор данных в соответствии с количеством переменных.

   б) Что выступает в качестве элементарной единицы?

   в) Это временной ряд или данные об одном временном срезе?

   г) Определите тип каждой переменной.

   д) Укажите, какие операции можно применять к каждой из переменных.

   е) Сформулируйте (в общих терминах), на какие вопросы, связанные с бизнесом, можно найти ответы при детальном анализе набора данных такого типа.

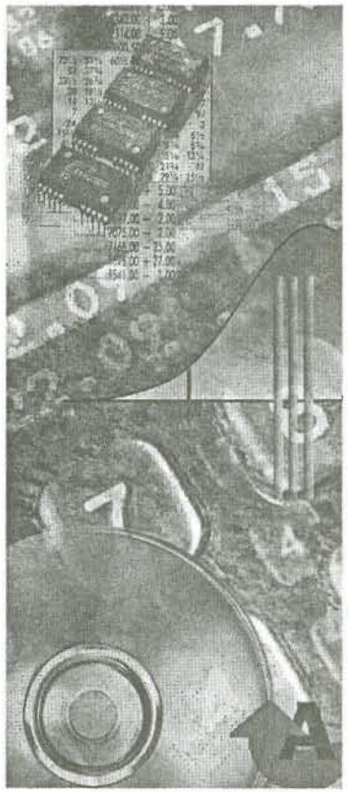
2. Воспользуйтесь таблицей данных из внутреннего отчета вашей фирмы или данными из вашего собственного опыта. Для этой таблицы дайте ответы на вопросы из п. 1.

3. Поищите в Internet данные об инвестициях в интересующую вас фирму. Кратко напишите о том, какая информация является доступной, и приложите распечатку некоторых числовых данных.

ГЛАВА 3

Гистограммы: взгляд на распределение данных

Ваш партнер уже полчаса рассматривает огромную таблицу расходов потребителей на покупку изделий ваших конкурентов, надеясь узнать как можно больше из чисел в колонках и даже отчасти преуспев в этом (об этом свидетельствуют периодические восклицания типа "Большинство тратит от 10 до 15 долларов!", "Практически никто не тратит больше 35 долларов!" и "О-о! Один потратил 58 долларов!"). Вы понимаете, что следует посоветовать партнеру использовать вместо этой таблицы какой-нибудь график, например гистограмму, поскольку это сэкономит время и даст более полную картину. Единственная проблема — чисто психологическая: как объяснить это партнеру, не задев его самолюбия.



В этой главе вы узнаете, как придать смысл колонке чисел. *Гистограмма* — это графическое изображение данных, которое дает визуальное представление многих основных свойствах набора данных в целом и позволяет ответить на следующие вопросы:

Первый. Какие значения типичны для этого набора данных?

Второй. Как различаются между собой значения?

Третий. Сконцентрированы ли данные вокруг некоторого типичного значения?

Четвертый. Какой характер имеет эта концентрация данных? В частности, одинаков ли характер "затухания" для малых и больших значений данных?

Пятый. Есть ли в этом наборе такие значения, которые настолько сильно отличаются от остальных, что требуют специальной обработки?

Шестой. Можно ли сказать, что это в целом однородный набор или отчетливо наблюдается наличие групп, которые необходимо анализировать отдельно?

Многие стандартные методы статистического анализа требуют, чтобы набор данных был приблизительно *нормально распределенным*. Вы узнаете, как распознать эту, похожую на колокол, форму и как преобразовать данные, если они не удовлетворяют этому требованию.

3.1. Последовательность данных

Набор данных простейшего вида — это **последовательность чисел**, представляющих некоторое свойство (единственная статистическая переменная), измеренное для каждого из рассматриваемых объектов (для каждой элементарной единицы). Последовательность чисел можно представить в нескольких, на первый взгляд сильно различающихся, формах. Помочь отличить результаты измерений (значения) от частот может ответ на вопрос: "Что представляют собой элементарные единицы, для которых проводились измерения?"

Пример. *Деятельность региональных менеджеров по продажам*

Рассмотрим пример очень короткой последовательности (только три наблюдения), где переменной является "объем продаж последнего квартала", а элементарными единицами — "региональные менеджеры по продажам".

Имя	Объем продаж (десятки тысяч)
Билл	28
Дженифер	32
Генри	18

Этот набор данных в дополнение к трем числам объема продаж содержит информацию для интерпретации (т.е. имя менеджера по продажам, которое помечает каждую элементарную единицу набора данных). Иногда такая первая колонка опускается, и значения переменной записываются непосредственно в первую колонку.

Пример. *Размер домохозяйства*

Иногда последовательность чисел имеет вид таблицы частот, как в приведенном ниже примере данных о количестве членов семьи в выборке из 17 домохозяйств.

Размер домохозяйства (количество человек)	Число домохозяйств (частота)
1	3
2	5
3	6
4	2
5	0
6	1

При интерпретации такой таблицы необходимо учитывать, что она представляет собой такую последовательность чисел, в которой каждое число из левой колонки (размер домохозяйства) повторяется такое количество раз, как указано в соответствующей строке в правой колонке (частота такого наблюдения,

количество домохозяйств такого размера). В таблице представлен следующий перечень чисел, отражающий количество людей в каждом домохозяйстве:

1, 1. 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 6

Число 1 повторяется в этом списке трижды (как показано в первой строке таблицы), число 2 — 5 раз (что следует из второй строки) и т. д.

Таблица частот особенно полезна для представления длинных перечней чисел с относительно небольшим количеством различных значений. Поэтому для выборки большого размера размеры домохозяйств можно было бы обобщить следующим образом.

Размер домохозяйства (количество человек)	Число домохозяйств (частота)
1	342
2	581
3	847
4	265
5	23
6	11
7	2

В этой таблице представлено много данных! Соответствующий перечень чисел начинается последовательностью из 342 единиц, затем идет 581 двойка и т.д. Таблица содержит размеры всех 2071 домохозяйства из этой большой выборки[1].

Числовая ось

Чтобы наглядно представить значения последовательности, мы расположим числа вдоль прямой. **Числовая ось** представляет собой прямую линию с нанесенной на ней шкалой числовых значений.



Важно расположить числа на оси равномерно и без пропусков[2]. Чтобы показать место каждого из чисел последовательности, можно сделать пометки в соответствующих местах на оси. Например, три цифры, соответствующие продажам:

28, 32, 18,

можно изобразить на числовой оси следующим образом.



Эта диаграмма дает наглядное представление о том, как эти значения соотносятся между собой. В частности, сразу же видно, что два значения относительно близки по величине между собой и намного больше третьего значения.

[1] Число 2071 — это итоговая частота, т.е. сумма всех чисел в правой колонке.

[2] Если необходимо разорвать числовую ось, например, чтобы пропустить не интересующие вас значения, то следует явно показать разрыв на числовой оси, что позволит не создавать ложное впечатление обычной непрерывной линии.

Использование такого рода или других графиков более информативно для анализа, чем просто рассматривание последовательностей чисел. Хотя числа хорошо подходят для регистрации данных, они не имеют наглядности в представлении целого ряда свойств данных. Например, последовательность

0 1 2 3 4 5 6 7 8 9

не дает никакого конкретного *визуального* указания о последовательном увеличении значений; при движении по перечню чисел слева направо числа не становятся больше по размеру, не становятся темнее и т.д. В то же время числовая ось явно показывает это важное свойство.

3.2. Использование гистограмм для отображения частот

Гистограмма демонстрирует частоты в виде диаграммы из столбиков, которые расположены над числовой осью и показывают, насколько часто различные значения встречаются в наборе данных. По горизонтальной оси откладывают измеренные значения из набора данных (выраженные в долларах, количестве людей, милях на галлон и других единицах измерения), по вертикальной — частоту встречаемости этих значений. Высоты прямоугольников соответствуют частотам значений, самый высокий столбик соответствует наиболее часто встречающемуся значению из набора данных, а самый низкий — значению, которое встречается реже всех.

Пример. *Процентные ставки ссуды под залог недвижимости*

В табл. 3.2.1 представлены размеры фиксированной процентной ставки ссуд под залог недвижимости, предоставляемых на 30 лет ипотечными компаниями Сиэтла.

Таблица 3.2.1. Ставки на ссуду под залог недвижимости

Кредитор	Процентная ставка
Accubanc Mortgage Corp.	7,000
Alpine Mortgage Services	6,875
American Investment Mrtg.	6,875
Bay Mortgage	6,750
Capital Mortgage Corp.	6,875
Castle Mortgage Corp.	7,250
Choice Mortgage	6,875
Citizen's Mortgage Inc.	7,000
City Mortgage	6,875
Community National Mrtg.	7,000
Countrywide Home Loans	7,250

Окончание табл. 3.2.1

Кредитор	Процентная ставка
Edmonds Mortgage Inc.	7,000
Equity Northwest, Inc.	7,000
Evergreen Pacific Services	6,125
First American Mortgage	6,750
First Mark Mortgage	7,125
First National Home Mrtg.	7,125
Goldmark Financial Corp.	7,000
Group One Mortgage, Inc.	7,000
Guaranty Mortgage Co.	7,000
Home Loans Online	6,875
Home Mortgage Corp.	6,875
Integral Real Estate&Mrtg	6,500
Intercontinental Mtg.	6,500
Lincoln Federal Mortgage	6,500
Merrill Lynch Credit	7,250
Millennium Mortgage	6,750
Mortgage Broker Services	6,875
Mortgage Network, Inc.	6,875
Mortgage Solutions	6,875
Nu-West Mortgage	6,875
Pacific Mountain Mortgage	6,500
Park Place	6,875
Performance Mortgage	7,000
Portia Financial Services	6,875
Prime Port Mortgage	7,000
Producer $ Mortgage Service	7,250
Raintree Financial Network	7,000
Redmond Mortgage Co.	6,625
Residential Mtg. Brokers Inc,	6,875
Select Mortgage	6,625
U.S. Discount Mortgage Co.	6,625
Wash. Women's Mrtg. Crp.	6,250
Western Heritage Mrtg. Svgs.	5,875
Wohletz Mortgage	7,000

Данные взяты из *The Seattle Times.* Spring Mortgage Rates, 1998, April, 19, p. C1.

Гистограмма показана на рис. 3.2.1

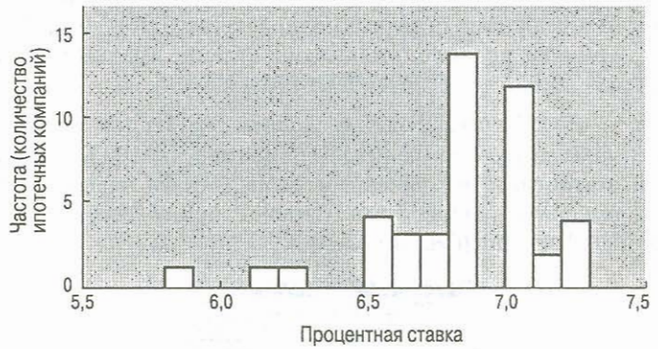


Рис. 3.2.1 Гистограмма процентных ставок ссуд под залог недвижимости

Теперь опишем общий подход к интерпретации гистограмм и одновременно выясним, что говорит нам о рассматриваемых процентных ставках этот конкретный график.

Числа на горизонтальной оси в нижней части рисунка указывают на значения процентных ставок, выраженные в процентах. Числа на вертикальной оси показывают частоту встречаемости каждой процентной ставки. Например, предпоследний столбик справа (расположенный по горизонтали между процентными ставками, равными 7,1% и 7,2%) имеет частоту (высоту), равную 2, означающую, что 2 финансовые организации предлагают ставку между 7,1% и 7,2%[3]. Таким образом, вы имеете графическое изображение характера изменения процентных ставок, которое показывает, какие значения встречаются наиболее часто, какие — наименее часто, а какие ставки вообще не предлагаются.

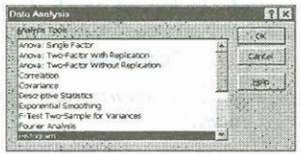
Что можно узнать о процентных ставках из этой гистограммы?

1. Размах (диапазон) значений. Размах процентных ставок составляет больше одной процентной единицы от наименьшего значения (около 5,8%) до наибольшего значения (около 7,3%) — это соответственно левая и правая границы гистограммы.
2. Типичные значения. Ставки размером от 6,8% до 7,1% встречаются чаще всего (обратите внимание на высокие столбики в этой части диаграммы).
3. Рассеяние. Типичная разница ставок для различных финансовых организаций составляет приблизительно 0,5% (умеренно высокие столбики отстоят друг от друга по горизонтальной оси приблизительно на 0,5 процентных единиц).
4. Общая конфигурация данных. Большинство организаций сосредоточены правее середины диапазона (здесь самые высокие столбики), и немного организаций предлагают либо очень низкие, либо очень высокие ставки (короткие столбики справа и слева).

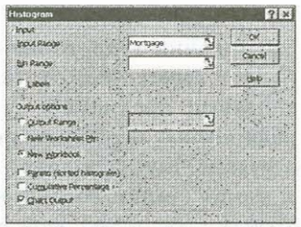
[3] Принято относить все значения данных, которые попадают точно между границами двух столбиков гистограммы, к столбику, расположенному справа. В данном конкретном случае столбик между числами 7,1 и 7,2 на горизонтальной оси включает все компании, чьи ставки равны или больше левого значения (7,1%), но меньше правого (7,2%). Организация, предлагающая 7,2%, войдет в следующий столбик, расположенный справа от значения 7,2.

5. Характерные особенности. Вероятно, вы заметили, что на гистограмме в этом примере пропущена область от 6,9 до 7,0. По-видимому, ни одна компания не предлагает ставку в интервале от 6,9% до 7,0%. Это обусловлено тем, что, как правило, указывают ставки, кратные 1/8 процента (например, 6,5%; 6,625%;, 6,75%, 6,875% и 7%).

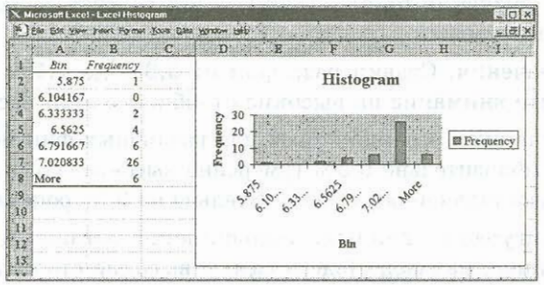
Несмотря на то что Microsoft Excel позволяет строить гистограммы, часто лучше использовать либо другое приложение (например, StatPad, см. приложение Г), либо отдельный статистический пакет программ. Чтобы построить гистограмму с помощью Excel, выберите в меню Tools (Сервис)[4] пункт Data Analysis (Анализ данных), а затем пункт Histogram (Гистограмма).



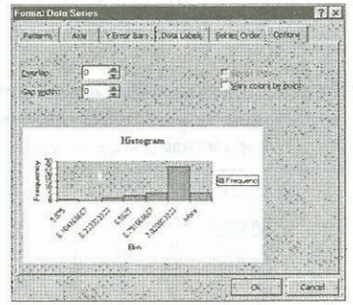
В появившемся диалоговом окне укажите данные (выделяя данные с помощью мыши в окне или, если данные названы, вводя соответствующее название), поставьте отметку возле Chart Output (Вывод графика) и укажите, куда необходимо произвести вывод.



Построенную гистограмму можно увеличить (щелкнуть мышью возле края и затем растянуть картинку), чтобы более отчетливо увидеть детали.



Однако столбики на этом рисунке слишком узкие, чтобы это была действительно гистограмма. Чтобы поправить это, дважды щелкните на изображении столбика, выберите в появившемся окне вкладку Options (Параметры), установите нулевое значение для параметра Gap Width (Ширина зазора) и щелкните на кнопке OK. В результате получим такую гистограмму.



Вы видите, что построить гистограмму в Excel непросто, особенно если вы хотите построить нестандартную гистограмму со столбиками определенной ширины (указывая значение для параметра Bin Range (Интервал карманов) в диалоговом окне). Поэтому в качестве альтернативы можно использовать StatPad (дополнение к Excel, см. приложение Г) или другой программный продукт.

Гистограммы и столбиковые диаграммы

Гистограмма — это столбиковая диаграмма частот, а не данных. Высота каждого столбика на гистограмме показывает, как часто указанное на горизонтальной оси значение встречается в наборе данных. Это дает визуальное представление о местах повышенной и пониженной концентрации данных. Каждый столбик на гистограмме может представлять много значений данных (фактически высота столбика точно отражает количество значений набора данных, которые принадлежат соответствующему диапазону). Это отличает гистограмму от столбиковой диаграммы фактических значений данных, где каждому определенному значению соответствует свой столбик. Также обратите внимание, что у гистограммы числа на горизонтальной оси всегда имеют содержательную интерпретацию, а у столбиковой диаграммы — не обязательно.

Пример. *Стартовая заработная плата выпускников USC с дипломами MBA*

Рассмотрим размер типичной начальной заработной платы в разных областях промышленности выпускников Южно-Калифорнийского университета (USC), получивших 1996 году степень магистра управления бизнесом (MBA). Соответствующие данные приведены в табл. 3.2.2.

Сравните гистограмму значений данных (рис. 3.2.2) и столбиковую диаграмму, приведенную на рис. 3.2.3. Обратите внимание, что столбики на гистограмме показывают количество отраслей в каждом из диапазонов заработной платы, а столбики на столбиковой диаграмме — фактическое значение заработной платы в конкретной отрасли.

[4] Если в меню Tools (Сервис) отсутствует пункт Data Analysis (Анализ данных), то сначала убедитесь, что вы выбрали ячейку электронной таблицы (а не график, например). Если вы все же не можете найти Data Analysis (Анализ данных), поищите пункт меню Add-Ins (Надстройки) и поставьте отметку возле Analysis ToolPak (Пакет анализа). Если это не поможет, то, видимо, необходимо переустановить Excel.

Полезны оба графических изображения. Столбиковую диаграмму лучше использовать, когда желательно идентифицировать все значения из набора данных, при условии, что набор данных достаточно небольшой. Однако для получения общего представления о наборе данных больше подходит гистограмма, особенно при больших наборах данных с множеством чисел.

Таблица 3.2.2. Стартовая заработная плата выпускников с дипломами MBA

Отрасль	Заработная плата, дол.
Аэрокосмическая	62 500
Автомобильная	50 000
Банковское дело	58 611
Компьютеры	59 280
Консультации	61 625
Потребительские товары	59 062
Электроника	58 016
Энергетика	63 333
Индустрия развлечений	55 000
Финансовые услуги	60 125
Инвестиционное банковское дело	58 500
Недвижимость	60 250
Розничная торговля	93 300

Данные представляют собой значения медианы заработной платы выпускников 1996 года Школы бизнеса Маршалла Южно-Калифорнийского университета, `http://www.usc.edu/dept/sba/career/emplorep.htm#class96`, 1998, November 2.

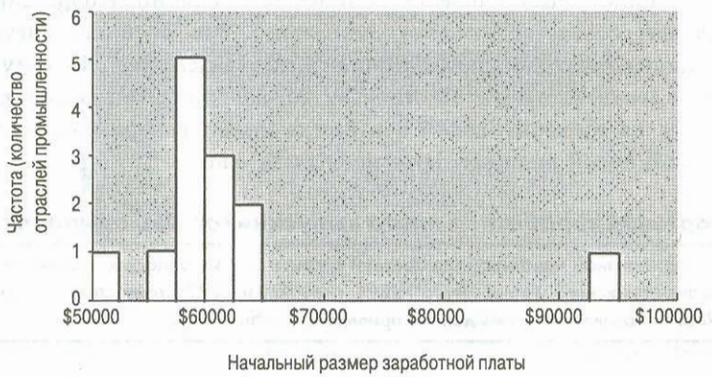


Рис. 3.2.2. Гистограмма значений начального размера заработной платы. Обратите внимание, что каждый столбик может представлять больше одной отрасли (см. число на вертикальной оси слева). Столбики показывают, какие диапазоны заработной платы чаще, а какие реже встречаются в этом наборе данных

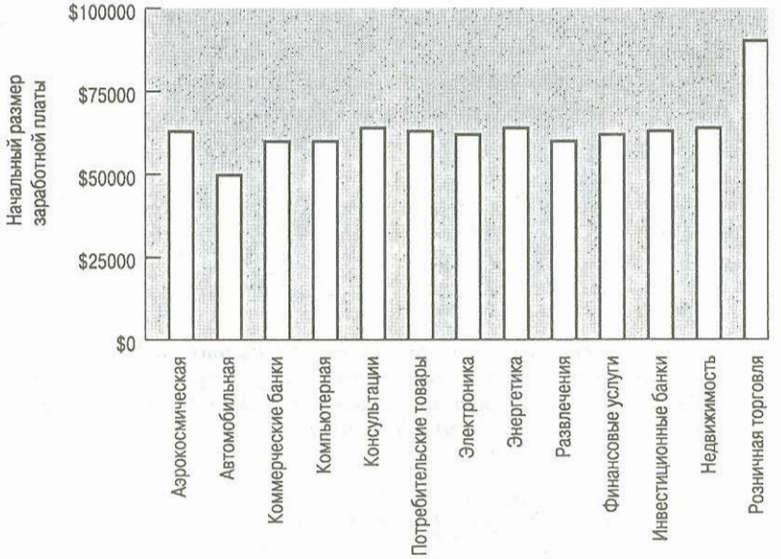


Рис. 3.2.3. Столбиковая диаграмма значений начального размера заработной платы. Обратите внимание, что каждый столбик представляет одну отрасль промышленности

3.3. Нормальное распределение

Нормальное распределение представляет собой теоретическую гладкую гистограмму в форме колокола без случайных отклонений. Такая кривая представляет идеальный набор данных, в котором большинство чисел сконцентрировано в средней части диапазона значений, а оставшиеся значения с затуханием симметрично расположены по обе стороны от вершины колокола. Такая степень гладкости не присуща реальным данным. На рис. 3.3.1 приведена кривая нормального распределения[5].

Фактически существует много различных кривых нормального распределения, форма которых напоминает симметричный колокол. Они отличаются расположением центра и масштабом (шириной колокола)[6]. Чтобы построить конкретную кривую нормального распределения, следует взять базовую кривую в форме колокола, переместить ее по горизонтали в точку, где предполагается разместить центр, а затем растянуть (или сжать). На рис. 3.3.2 приведено несколько кривых нормального распределения.

Почему нормальное распределение играет такую важную роль в статистике? Обычно в статистике предполагают, что распределение данных приблизительно

[5] Для любознательных приведу формулу этой колоколоподобной кривой: $\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)/\sigma^2/2}$, где μ — центр, описанный в главе 4, определяет горизонтальное положение наивысшей точки, а σ определяет ширину колокола (изменчивость или масштаб, о которых речь идет в главе 5).

[6] Эти понятия будут подробно рассмотрены в главах 4 и 5.

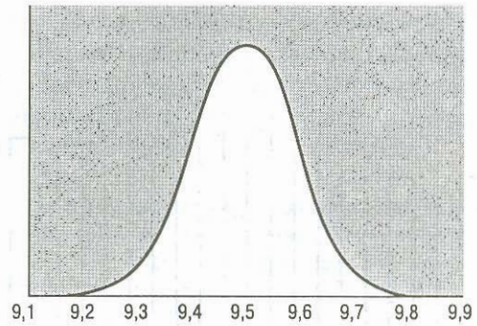


Рис. 3.3.1. Идеальная (теоретическая) кривая нормального распределения. Реальные нормально распределенные наборы данных имеют некоторые случайные отклонения от этой идеально гладкой кривой

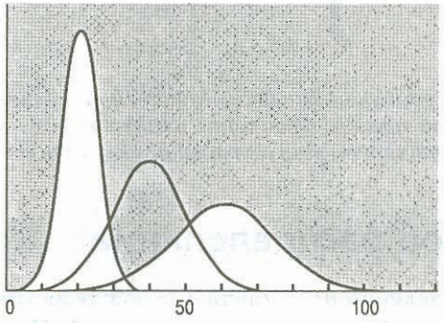


Рис. 3.3.2. Кривые нормального распределения с различными центрами и по-разному растянутые (в разных масштабах)

соответствует нормальному[7]. Специалисты-статистики знают свойства нормального распределения и используют их всякий раз, когда гистограмма похожа на кривую нормального распределения.

В каком случае можно сказать, что набор данных подчиняется нормальному распределению? Хороший способ заключается в том, чтобы внимательно изучить гистограмму. На рис. 3.3.3 представлены гистограммы для различных выборок объемом 100 значений каждая из нормально распределенного набора данных. Этот рисунок демонстрирует, насколько случайной может быть форма распределения при ограниченном размере выборки.

Уменьшение количества данных приводит к увеличению случайности, поскольку нет достаточно информации для представления полной картины распределения. Это наглядно показано на рис. 3.3.4, где приведены гистограммы для выборок объемом 20 значений каждая из нормально распределенного набора данных.

[7] В частности, многие стандартные методы для вычисления доверительных интервалов и проверки статистических гипотез (о которых вы узнаете позже) требуют, чтобы данные были распределены нормально (по крайней мере, приблизительно).

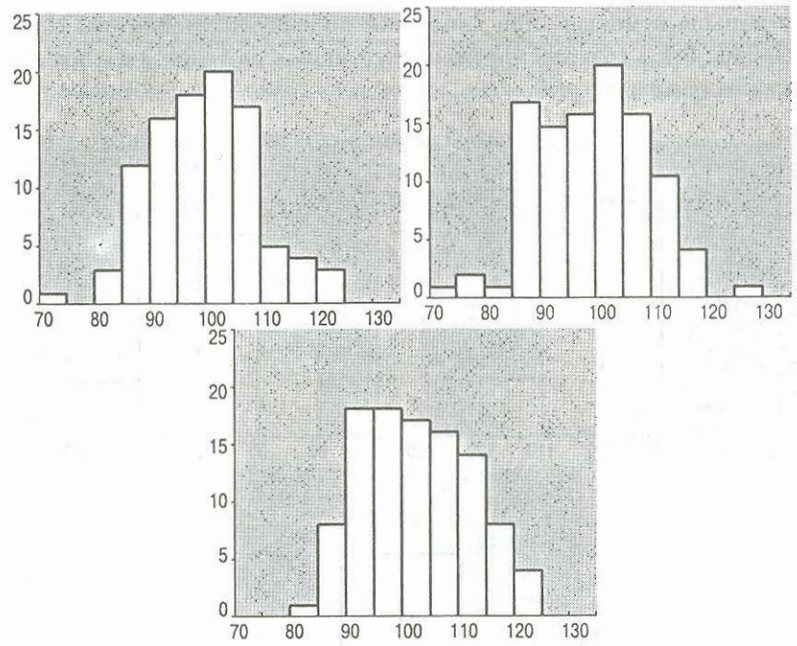


Рис. 3.3.3. Гистограммы для данных, извлеченных из нормально распределенного набора. Объем каждой выборки равен 100. Сравнение этих трех гистограмм демонстрирует, какую степень случайности можно ожидать

Действительно ли в реальной жизни все наборы данных подчиняются нормальному распределению? Конечно, нет. Используя гистограмму, важно определить, являются ли данные нормально распределенными. Это особенно важно, если дальнейший анализ предполагает использование стандартных статистических процедур, которые требуют нормального распределения данных. В следующем разделе мы рассмотрим один вид отличия экономических данных от нормального распределения и предложим способ справиться с этой проблемой.

3.4. Несимметричные распределения и преобразование данных

Несимметричное (скошенное) распределение не является ни симметричным, ни нормальным, поскольку значения данных на одной стороне кривой затухают быстрее, чем на другой. В бизнесе часто можно встретить асимметрию в наборах данных, которые отражают величины, выраженные положительными числами (например, объемы продаж или размеры активов). Это связано с тем, что такие данные не могут принимать отрицательные значения (наличие границы с одной стороны) и значения не ограничены сверху. В результате на гистограмме много значений данных сконцентрировано около нуля, и количество значений стано-

вится все меньше и меньше при движении по горизонтальной оси гистограммы вправо. На рис. 3.4.1 содержится несколько примеров теоретических кривых несимметричных распределений.

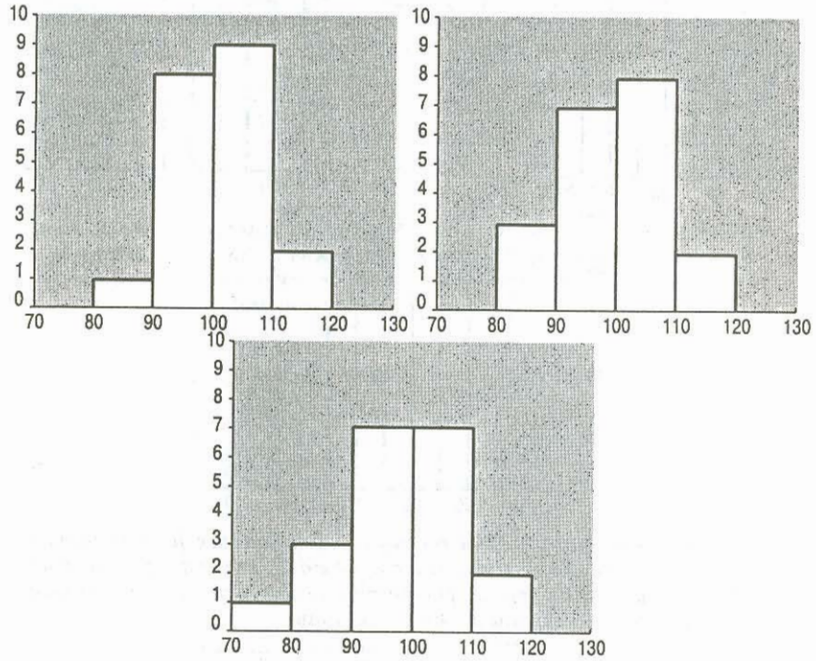


Рис. 3.3.4. Гистограммы для данных, извлеченных из нормально распределенного набора. Объем каждой выборки равен 20. Сравнение этих трех гистограмм демонстрирует, какую степень случайности можно ожидать

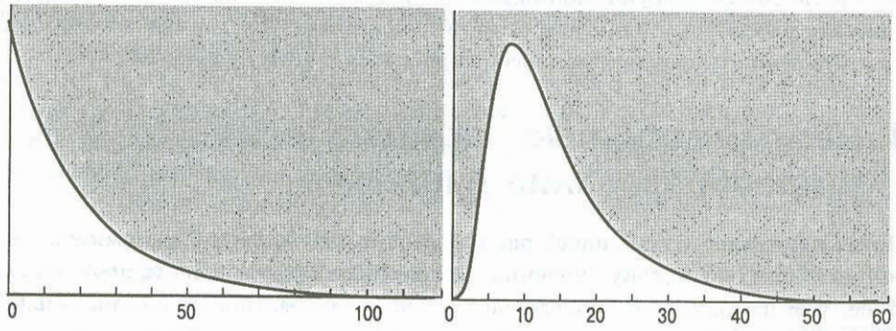


Рис. 3.4.1. Примеры сглаженных идеализированных кривых несимметричных распределений. Реальные асимметрично распределенные наборы данных имеют некоторые случайные отклонения от таких идеально гладких кривых

Пример. *Активы коммерческих банков*

В табл. 3.4.1 содержатся данные об активах коммерческих банков из списка Fortune 1000. Эти данные представляют хороший пример сильно несимметричного (сильно скошенного) распределения.

Таблица 3.4.1. Активы коммерческих банков из Fortune 1000

Банк	Активы, млрд дол.
Chase Manhattan Corp.	366
Citicorp	311
NationsBank Corp.	265
J. P. Morgan & Co	262
BankAmerica Corp.	260
First Union Corp.	157
Bankers Trust New York Corp.	140
Bank One Corp.	116
First Chicago NBD Corp.	114
Wells Fargo & Co.	97
Norwest Corp.	89
Fleet Financial Group	86
PNC Bank Corp.	75
KeyCorp	74
U. S. Bancorp	71
BankBoston Corp.	69
Wachovia Corp.	65
Bank of New York Co.	60
Sun Trust Banks	58
Republic New York Corp.	56
National City Corp.	55
CoreStates Financial Corp.	48
Barnett Banks	47
Mellon Bank Corp.	45
State Street Corp.	38
Comerica	36
South Trust Corp.	31
Summit Bancorp	30
Mercantile Bancorp.	30
BB&T Corp.	29
Huntington Bancshares	27
Northern Trust Corp.	25

Окончание табл. 3.4.1

Банк	Активы, млрд дол.
Crestar Financial Corp.	23
Regions Financial	23
Fifth Third Bancorp	21
MBNA	21
First of America Bank Corp.	21
Firstar Corp.	20
Marshall & Ilsley Corp.	19
Popular	19
Amsouth Planters Corp.	18
First Security Corp.	17
Pacific Century Financial	15
First Tennessee National Corp.	14
First Empire State Corp.	14
Old Kent Financial Corp.	14
Compass Bancshares	13
Synovus Financial Corp.	9
First National of Nebraska	7
Providian Financial	4

Данные взяты из списка коммерческих банков Fortune 500 Industry с использованием ссылки `http://www.pathfinder.com/fortune500/ind30.html`, 1998, November 2.



На рис. 3.4.2 приведена гистограмма этого набора данных. Из-за асимметрии распределение данных нельзя отнести к нормальному.

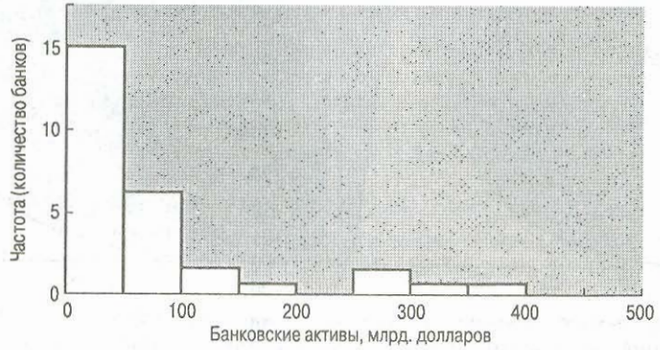


Рис. 3.4.2. Гистограмма размеров активов (в миллиардах долларов) коммерческих банков по данным Fortune 1000. Это асимметричное (скошенное), а не нормальное распределение

Очень высокий столбик слева на гистограмме представляет большинство из этих банков, которые имеют активы менее 50 миллиардов долларов. Несколько столбиков, расположенных правее, представляют относительно небольшое число более крупных банков. Очень небольшой столбик справа на гистограмме представляет один банк — Chase Manhattan Corp., который имеет активы 366 миллиардов долларов.

Пример. *Численность населения штатов*

Другой пример несимметричного распределения дают данные о численности населения отдельных штатов в США, представленные в виде списка чисел[8]. Асимметрия отражает тот факт, что есть много штатов с небольшим или средним размером населения и всего несколько штатов с большим размером населения (три самых больших штата: Калифорния, Техас и Нью-Йорк). Гистограмма приведена на рис. 3.4.3.

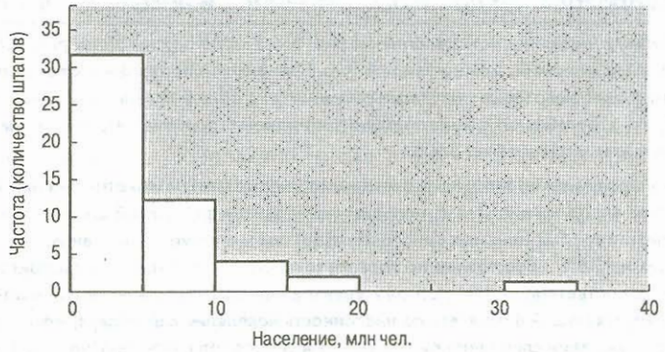


Рис. 3.4.3. Гистограмма численности населения штатов в 1996 году: несимметричное распределение

Проблема с асимметрией

Одна из проблем, связанных с асимметрией данных, состоит в том, что многие из наиболее распространенных статистических методов (о которых вы узнаете в следующих главах) требуют, чтобы данные были, по крайней мере, приблизительно нормально распределенными. Если эти методы применяют к несимметричным данным, то полученный результат может быть неточным или просто неверным. И даже если результаты получаются в основном корректными, будет определенная потеря эффективности анализа, поскольку не обеспечивается наилучшее использование всей информации, содержащейся в наборе данных.

Выход с помощью преобразования

Один из способов справиться с проблемой асимметрии заключается в использовании такого *преобразования*, которое переводит несимметричное распределение в более симметричное. **Преобразование** заключается в замене каждого значения набора данных другим числом (например, логарифмом этого значения) с целью упростить статистический анализ. Наиболее распространенным типом преобразования данных в бизнесе и экономике является логарифмирование, которое можно использовать только для положительных чисел (т.е. для данных,

[8] *Statistical Abstract of the United States*: 1997, Washington, Dc, 1997, p. 28.

которые включают отрицательные значения или нуль, этот метод не подходит). **Логарифмирование** часто преобразует скошенные (асимметричные) данные в симметричные, поскольку происходит растягивание шкалы возле нуля, что, в свою очередь, приводит к распределению малых значений, сгруппированных вместе. В то же время логарифмирование собирает вместе большие значения, которые распределены на правом (положительном) конце шкалы. Оба типа наиболее часто используемых логарифмов ("десятичный логарифм" по основанию 10 и "натуральный логарифм" по основанию "*e*") одинаково эффективно можно использовать для такого рода преобразований. В этом разделе мы будем использовать десятичный логарифм.

Пример. *Преобразование данных о численности населения штатов*

Сравнивая гистограмму численности населения на рис. 3.4.3 с гистограммой на рис. 3.4.4, построенной для логарифмов тех же значений, можно увидеть, что в результате логарифмирования асимметрия исчезает. Хотя и в этом случае некоторые результаты выпадают из общей картины распределения и кривая не идеально симметрична, больше нет резкого падения на одной стороне и медленного уменьшения значений на другой, как имеет место на рис. 3.4.3.

Логарифмическую шкалу можно интерпретировать скорее как *мультипликативную* или *процентную*, чем как аддитивную. Как видно на рис. 3.4.4, использование логарифмической шкалы приводит к тому, что расстояние по горизонтали 0,2 (ширина одного столбца) соответствует увеличению (при движении слева направо) населения на 58%[9]. Расстояние по горизонтальной оси, равное пяти столбикам (например, от точки 6 до точки 7), соответствует 10-кратному увеличению численности населения штата[10]. На первоначальной шкале (отражающей фактическую численность населения в штатах) трудно проводить сравнение процентов. При движении слева направо на рис. 3.4.3 переход к каждому новому столбику означает увеличение населения на 5 миллионов человек, но в левой части рисунка эта разница составляет значительно больший процент, чем в правой.

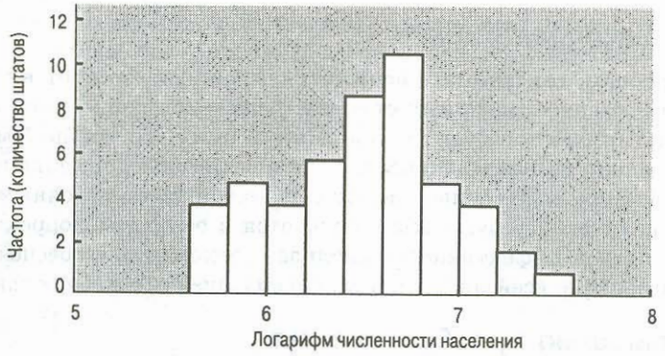


Рис. 3.4.4. Преобразование может позволить перейти от асимметрии к симметрии. Гистограмма логарифмов значений численности населения штатов за 1996 год в основном симметрична, за исключением случайных отклонений. По существу, никакой систематической асимметрии не осталось

[9] Поскольку $10^{0,2} = 1,58$, что на 58% больше, чем 1.

[10] Поскольку $10^1 = 10$.

Интерпретация и вычисление логарифма

Разница на 1 для значений логарифма по основанию 10 соответствует десятикратной разнице для исходных значений. Например, значения 392,1 и 3921 (соотношение 1:10, разница в 10 раз) после логарифмирования преобразуются соответственно в значения 2,59 и 3,59 (разница на 1). В табл. 3.4.2 содержатся примеры нескольких чисел и их логарифмов.

Из таблицы видно, как логарифм "стягивает вместе" очень большие числа, уменьшая разницу между ними и другими значениями в наборе данных (например, вместо разницы в 100 миллионов получаем разницу в 8 единиц). Кроме того, обратите внимание, что логарифм приблизительно показывает количество разрядов в целой части числа. Например, население Калифорнии составляет 31878234 человека, а логарифм этого числа равен 7,5035 (это соответствует столбикам в правой части рис. 3.4.3 и 3.4.4).

Чаще всего используют логарифмы двух видов. Мы рассмотрели логарифмы по основанию 10. Логарифмы второго вида называют натуральными, их обозначают ln и вычисляют по основанию числа e=2,71828.... Натуральный логарифм часто используют при вычислении сложных процентов, темпов роста, экономической эластичности и др. В преобразованиях данных оба вида логарифмов приводят к одинаковому эффекту, т.е. "стягивают вместе" на числовой оси большие числа и "растягивают" малые.

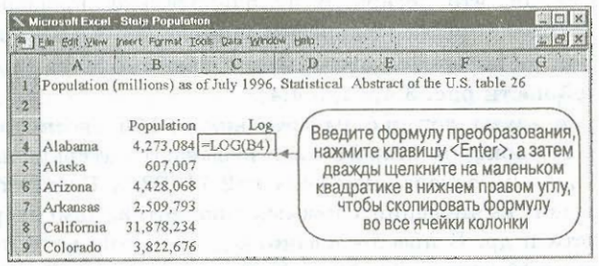
Таблица 3.4.2. Результаты логарифмирования по основанию 10

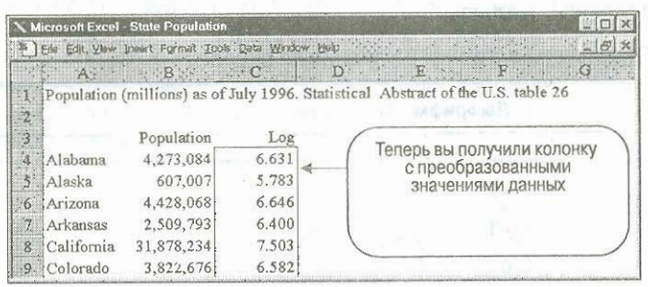
Число	Логарифм
0,001	–3
0,01	–2
0,1	–1
1	0
2	0,301
5	0,699
9	0,954
10	1
100	2
10 000	4
20 000	4,301
100 000 000	8

Логарифмировать можно на калькуляторе, используя клавишу LOG[11]. Необходимо просто ввести число и нажать клавишу LOG. Многие электронные таблицы, например Microsoft Excel, имеют встроенные функции логарифмирования. Вы

[11] В некоторых калькуляторах вместо функции LOG для вычисления десятичных логарифмов есть функция LN для вычисления натуральных логарифмов (по основанию *e*). Чтобы вычислить обычный десятичный логарифм на таком калькуляторе, разделите результат применения LN на число 2,302585 — значение натурального логарифма для числа 10.

можете ввести =LOG(5) в ячейку, чтобы вычислить логарифм (по основанию 10) числа 5, равный 0,69897. Если же вы введете =LN(5), то получите логарифм числа 5 по основанию *e*, равный 1,60944. Чтобы найти логарифм набора данных в колонке таблицы, можно с помощью команд Copy (Копировать) и Paste (Вставить) скопировать формулу логарифма из первой ячейки во все ячейки колонки, что существенно сокращает время вычисления логарифмов для ряда чисел. Еще более быстрый способ создания колонки преобразованных значений показан ниже: следует после ввода формулы преобразования дважды щелкнуть на "быстром заполнении" ("fill handle", маленький квадрат, расположенный справа ниже выделенной ячейки) или просто протащить "быстрое заполнение" мышью.





3.5. Бимодальные распределения

Важно уметь определять, когда набор данных состоит из двух или более отчетливо различающихся между собой групп, чтобы можно было при необходимости анализировать эти группы отдельно. На гистограмме такой ситуации соответствует разрыв между двумя соседними группами столбиков. Если на гистограмме четко видны две отдельные группы, то это говорит о **бимодальном распределении** данных. Бимодальное распределение — это распределение, имеющее *две моды* или два различных кластера (блока) данных[12].

Наличие бимодального распределения может свидетельствовать о том, что ситуация более сложная, чем вы предполагали, и поэтому требует серьезного внимания. По меньшей мере, следует выявить причины наличия двух групп. Возможно, интерес представляет только одна группа, поэтому другую группу можно исключить из рассмотрения. А может быть, вам необходимо изучить обе группы, но следует внести некоторые уточнения, чтобы учесть факт имеющегося различия.

[12] Статистический показатель *мода* будет описан в главе 4.

Пример. *Доходы валютного рынка*

Рассмотрим доходы валютного рынка как ежегодные дивиденды, выраженные в процентах. В табл. 3.5.1 представлена часть соответствующего набора данных.

Гистограмма полного набора данных, показанная на рис. 3.5.1, выглядит как две отдельные гистограммы. Одна группа содержит фонды с доходами от 3 до 4%, а другая — от 6 до 8%. Маловероятно, что такое разделение обусловлено простой случайностью для одного однородного набора данных. Видимо, существует другая причина (попробуйте угадать ее, прежде чем посмотрите ответ, представленный в ссылке)[13].

Таблица 3.5.1. Доходы взаимных фондов на валютном рынке (часть списка)

Фонд	Доход за 7 дней, %	Фонд	Доход за 7 дней, %
Putnam MMA	5,3	DryMATR	3,71
QualivestA	4,86	DryCTMu	3,42
QualivestGvA	4,88	DreyCalTx	3,31
QualivestY	5,25	DrMI Mun	3,55
QltvCsh	4,81	DrNJMun	3.43
QuestCshGov	4,75	DrNYTE	3,44
QuestCshPr	4,85	DRPA Mun	3,70
RNC Liq	4,82	Drey TxEx C	3,63
RegisDSI	5,61	DryMuResR	3,44
RemTreasTr	5,01	DryMAMun	3,39
RemGovTr	5,37	DryOHMu	3,76
RembMMTr	5,40	DryCATF	3,44
ResrevUSTrs	4,64	DryMATF	3,46
ResrveFd Gvt	4,99	EatnVn	3,44
ReserveFd	5,01	Evrgn TEA	3,71
:	:	:	:

Данные взяты из *Wall Street Journal*, 1995, October, 5, с. C22, колонки 5 и 6 из 8. В представленных там таблицах четко выделено (с помощью заголовка "Tax Exempt" – "Свободный от налога"), где заканчиваются налогооблагаемые фонды и начинаются фонды, освобожденные от уплаты налога.

[13] Исходный набор данных содержит заголовок "Свободный от налога", который отделяет в списке последний из обычных налогооблагаемых фондов от тех, что вкладывают средства только в необлагаемые налогом ценные бумаги. Поскольку для необлагаемых налогом фондов не начисляется налог от полученного процента, то эффективный доход (с поправкой на налог) выше, чем то значение, которое обычно указывают. Таким образом, группа с более низкими доходами в левой части гистограммы (рис. 3.5.1) включает фонды, освобожденные от уплаты налога. Если вы хотите обобщить текущие рыночные процентные ставки, то доходы фондов, освобожденных от налога, необходимо предварительно обработать. Можно не рассматривать необлагаемые налогом фонды и проанализировать только доходы налогооблагаемых фондов. С другой стороны, можно предварительно откорректировать необлагаемые налогом доходы, чтобы привести их в соответствие с остальными, и затем продолжить анализ всего откорректированного набора данных.

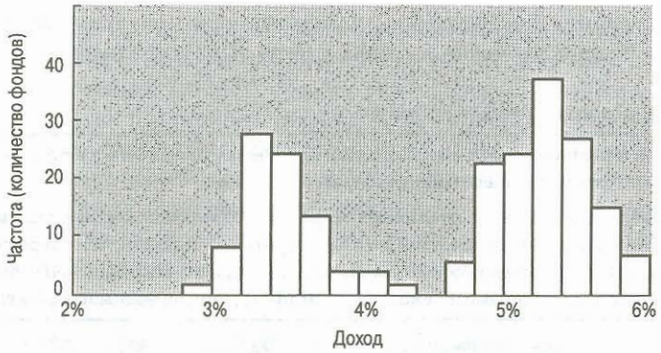


Рис. 3.5.1. Доходы валютного рынка. Это бимодальное распределение с двумя четкими отдельными группами, что, видимо, не может быть объяснено только случайностью

Пример. *Стоимость одного дня пребывания в больнице*

Рассмотрим стоимость одного дня пребывания в местной больнице в разных штатах (табл. 3.5.2).

Отметим значительное варьирование стоимости в разных штатах: стоимость одного дня пребывания в больнице Калифорнии почти вдвое выше ($1134), чем в больнице Южной Дакоты ($457). Чтобы представить полную картину, посмотрите на гистограмму этого набора данных на рис. 3.5.2.

Это почти симметричное распределение (т.е. оно не идеально симметричное, но по крайней мере не является сильно асимметричным). Распределение в основном нормальное, данные содержат только одну группу значений.

Однако если начертить ту же гистограмму в уменьшенном масштабе, с более узкими столбиками (рис. 3.5.3), то дополнительно в данных обнаружатся две группы значений: 5 штатов с более низкой стоимостью пребывания в больнице (слева) и остальные штаты (справа), а может быть, даже и три группы.

Однако это распределение не является действительно бимодальным по двум причинам. Во-первых, разрыв мал по сравнению с разбросом значений стоимости в разных штатах. Во-вторых, и это более важно, столбики гистограммы слишком малы, потому что многие из них представляют только один штат. Помните, что одна из основных целей статистических методов (таких как гистограмма) заключается в выявлении *общей картины*, а не отдельных деталей.

Таблица 3.5.2. Средняя стоимость одного дня пребывания в местной больнице одного пациента (в долларах)

Alabama	729	Kentucky	674	North Dakota	484
Alaska	1 116	Louisiana	836	Ohio	875
Arizona	1 051	Maine	674	Oklahoma	740
Arkansas	633	Maryland	806	Oregon	1 011
California	1 134	Massachusetts	937	Pennsylvania	793
Colorado	904	Michigan	847	Rhode Island	801
Connecticut	1 012	Minnesota	618	South Carolina	782
Delaware	920	Mississippi	516	South Dakota	457

Окончание табл. 3.5.2

Dist. of Columbia	1 124	Missouri	792	Tennessee	796
Florida	886	Montana	474	Texas	933
Georgia	721	Nebraska	600	Utah	1 036
Hawaii	761	Nevada	952	Vermont	726
Idano	618	New Hampshire	776	Virginia	774
Illinois	849	New Jersey	737	Washington	974
Indiana	822	New Mexico	950	West Virginia	655
Iowa	588	New York	744	Wisconsin	674
Kansas	661	North California	711	Wyoming	515

Данные представлены Бюро переписи населения США, *Statistical Abstract of the United States*: 1994, 114th ed. (Washington, D. C., 1994), таблица 183 содержит данные за 1992 год. Источниками данных являются ежегодный отчет (защищено авторским правом) Американской ассоциации больниц, Чикаго, *Hospital Statistics*, и неопубликованные данные.

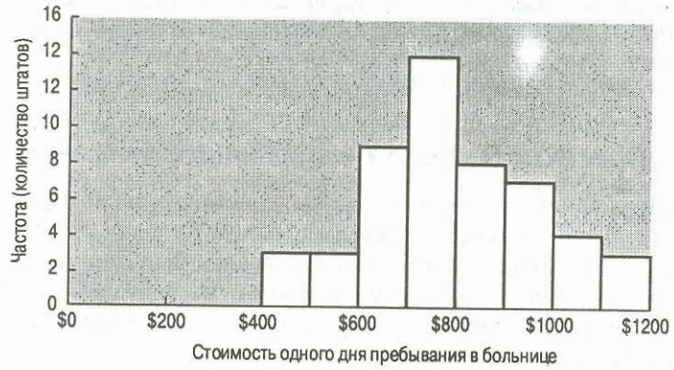


Рис. 3.5.2. Стоимость одного дня пребывания в муниципальной больнице в разных штатах. Это почти нормальное распределение, образующее одну целостную группу

3.6. Выбросы (сильно отклоняющиеся значения)

Иногда в данных можно наблюдать **выбросы (сильно отклоняющиеся значения)**, т.е. такие значения, которые, по-видимому, не принадлежат данному распределению, поскольку они либо слишком велики, либо слишком малы. В зависимости от причин, вызвавших выбросы, проблему выбросов решают по-разному. Существуют два вида выбросов значений: ошибки и корректные, но "отличающиеся" значения данных. Поскольку о выбросах часто говорят при анализе гистограмм, мы их обсудим в этой главе. А в следующей главе будет изложен формальный метод вычислений для определения выбросов (построение подробной блочной диаграммы).

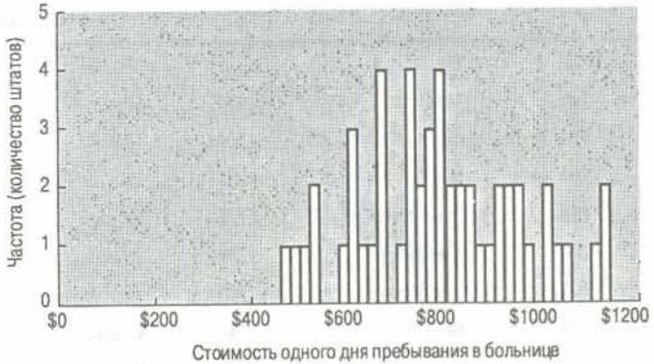


Рис. 3.5.3. Стоимость одного дня пребывания в муниципальной больнице в разных штатах (те же данные, что и на предыдущей гистограмме, но здесь используются более узкие столбики). Поскольку на этой гистограмме представлено больше деталей, создается впечатление (возможно, неверное), что в наборе данных присутствуют две или даже три группы. Пять штатов слева с наименьшей стоимостью немного отделены от остальных, как и три штата справа с наибольшей стоимостью. Возможно, это просто случайность и не является действительной бимодальностью

Работа с выбросами (сильно отклоняющимися значениями)

С ошибками справиться легко — нужно просто откорректировать значение. Например, если значение, соответствующее объему продаж $1597,00, записано как 159700 из-за неправильно поставленной десятичной точки, то это значение будет сильно отличаться от остальных значений на гистограмме. Увидев такое странное значение, нужно перепроверить данные и найти ошибку. Исправив это значение на 1597, вы решите проблему.

К сожалению, труднее решать проблемы выбросов корректных данных. Если есть убедительное подтверждение того, что выбросы не соответствуют тому, что изучается, то их можно просто удалить и анализировать оставшиеся более согласованные между собой данные. Например, в наборе данных относительно доходов денежного рынка может появиться несколько значений доходов фондов, не облагаемых налогом. Если цель исследования состоит в анализе рыночной ситуации для обычных фондов, то эти выбросы лучше исключить из общей картины. В качестве другого примера предположим, что ваша компания оценивает новый фармацевтический продукт. В одном из опытов лаборант чихнул в образец перед его анализом. Если вы не изучаете несчастные случаи с лабораторными материалами, то этот образец можно не анализировать.

Если вы решили не учитывать некоторые выбросы, вы должны быть готовы к тому, что в правильности этого решения нужно убедить не только себя, но и того, кому предназначен ваш отчет (хотя этот человек может иметь и другое мнение). Таким образом, на вопрос, учитывать или не учитывать выбросы, нет однозначного и единственно верного ответа. Например, для упрощения первоначального внутрифирменного анализа можно исключить некоторые выбросы. Однако

если исследование предназначено для общественности или представляет собой государственное исследование, то следует очень осторожно и со всей ответственностью отнестись к исключению выбросов значений.

При отсутствии достаточно обоснованного аргумента для исключения выбросов как компромисс можно выполнить *два различных анализа*: один с учетом выбросов, а другой — с исключением их. Тогда ваш отчет будет содержать все результаты. В лучшем случае результаты обоих анализов будут одинаковыми, тогда можно будет сделать вывод, что наличие выбросов не имеет существенного значения. В более сложном случае, когда эти два анализа дадут разные результаты, ваши выводы и рекомендации будут менее определенными и однозначными. К сожалению, нет исчерпывающего решения этой достаточно тонкой проблемы[14].

При исключении из анализа выбросов рекомендуется руководствоваться одним важным правилом, которое поможет вам защитить себя от возможных обвинений.

При исключении выбросов из анализа

Всегда объясняйте, что вы сделали и почему!

Другими словами, четко объясните в отчете (может быть, достаточно сноски), что ваши данные содержат выбросы (сильно отклоняющиеся значения). Опишите эти значения. Объясните и обоснуйте предпринятые вами действия.

Почему проблемы с выбросами нужно обязательно решать? Есть две причины, по которым наличие выбросов может приводить к проблемам при анализе данных. Во-первых, трудно интерпретировать подробности структуры набора данных, если одно значение доминирует в общей картине и поэтому привлекает к себе повышенное внимание. Во-вторых, как и в случае асимметрии, многие из распространенных современных статистических методов нельзя использовать для анализа тех данных, распределение которых сильно отличается от нормального. Нормальное распределение является симметричным и обычно не содержит выбросы. Следовательно, прежде чем заняться серьезными статистическими выводами, вам придется разобраться с выбросами в данных.

Пример. *Растут или падают чистые поступления?*

По сообщению *The Wall Street Journal*[15], чистый доход за второй квартал крупнейших компаний США возрос на 27% (по результатам анализа данных о 677 открытых акционерных торговых компаниях). Однако в данных есть выбросы значений: в результате отделения от компании U.S. West доход компании MediaOne составил во втором квартале 24,5 миллиардов долларов. Если это значение исключить из анализа, то увеличение чистого дохода фактически упадет до 1,5%.

Почти такая же ситуация наблюдалась в предыдущем квартале, когда чистый доход возрос на 20% благодаря продажам компании Ford Motors. Если исключить этот выброс, то вместо сильного роста получим просто рост на 2,5%.

---

[14] В современной статистике есть раздел "устойчивость" (робастность), в котором применяется мощный вычислительный аппарат для учета наличия выбросов значений, а также разрабатываются устойчивые методы, доступные для многих (но не для всех) наборов данных. Более подробно об этом см. в Hoaglin D. C., Mosteller F. and Tukey J. W. *Understanding Robust and Explanatory Data Analysis* (New York: Wiley, 1983); Barnett V. and Lewis T. *Outliers in Statistical Data* (New York: Wiley, 1978).

[15] Philips M.M. MediaOne Item Pushes Earnings of U.S. Firms to Gain, but Asia and Competition Hurt Results. *The Wall Street Journal*, 1998, August, 3, p. A1, C15.

Из этих примеров видно, что при наличии выброса значений статистическое обобщение результатов может быть ошибочным. Если вы прочитаете только эти показатели крупных компаний — увеличение на 27% (или 20%), — то можете сделать неверный вывод о том, что большинство компаний испытывают сильный экономический рост. Исключив выбросы и сделав повторный анализ, получим более реальное впечатление о ситуации в этой группе компаний.

Пример. *Изменения в расходах на телевизионную рекламу*

В рекламном бизнесе, как и в большинстве сфер экономической деятельности, время от времени происходят изменения. В табл. 3.6.1 приведены процентные изменения общих расходов на телевизионную рекламу в 1994 г. по сравнению с 1993 г. для 25 самых крупных рекламодателей 1994 г.

На гистограмме, показанной на рис. 3.6.1, наличие выброса (изменение расходов для Regal Communications представляет собой 2353,7% — колоссальное число, которое соответствует увеличению расходов с 1,1 до 25,8 миллиона долларов) привело к тому, что все остальные компании оказались сведены в один столбик, который соответствует объему увеличения расходов рекламодателей где-то между 0 и 500%. Исключением является одна компания, которая увеличила расходы более чем на 500%, и одна компания, которая уменьшила расходы.

Изменение расходов компании Regal Communications закрывает картину распределения процентных изменений расходов других компаний (большинство значений лежат в пределе от 0 до 100%). Даже построив гистограмму из узких столбиков (рис. 3.6.2), нельзя увидеть подробную картину. К сожалению, гистограмма в данном случае не очень полезна.

Таблица 3.6.1. Изменение общих расходов на телевизионную рекламу в 1994 году по сравнению с 1993 г.

Рекламодатель	Изменение расходов на телерекламу, %	Рекламодатель	Изменение расходов на телерекламу, %
Procter & Gamble	43,2	Warner-Lambert	–22,7
Phillip Morris	27,5	AT&T	73,5
Kellogg	77,9	Grand Metropolitan	14,0
Time Warner	201,0	Johnson & Johnson	16,5
Unilever	16,7	National Education	217,3
Hasbro	54,5	Nestle	31,4
Mattel	47,7	Hershey	42,4
American Home Products	104,4	Regal Communications	2 353,7
General Motors	65,7	McDonald's	28,5
Wrigley	66,8	Sara Lee	16,4
Mars	33,3	Himmel Group	684,0
RJR Nabisco	65,9	Bayer Group	12,7
Sears, Roebuck	44,7		

Источник: Top 25 syndicated TV advertisers, *Advertising Age*, 1995, September 27, p. 42. На основе информации из *Competitive Media Reporting*.

Исключение компании Regal Communications, которая, очевидно, представляет выброс данных с наибольшим значением (2353,7%), преследует цель получить более структурированную картину в отношении других рекламодателей. Однако, как видно из рис. 3.6.3, большинство деталей все еще скрыто, но на этот раз другим выбросом, равным 684% (компания Himmel Group)

После исключения обоих выбросов можно наконец увидеть, что распределение изменений расходов на телерекламу для оставшихся рекламодателей является приблизительно нормальным, с центром около 40% (рис. 3.6.4), но, возможно, с еще двумя сильно отклоняющимися значениями (выбросами) — около 200% (компании Time Warner и National Education)

На основе этого анализа изменений расходов на телевизионную рекламу можно дать следующую оценку ситуации.

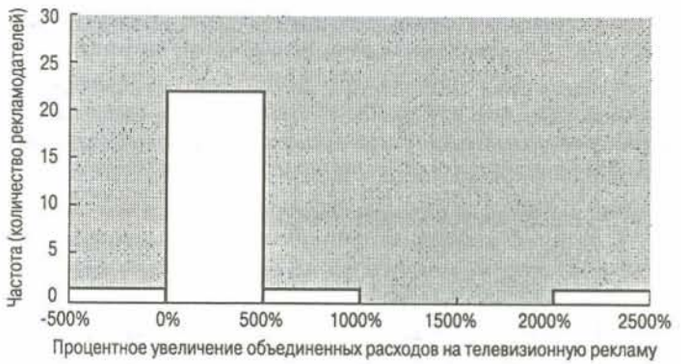


Рис. 3.6.1. Гистограмма процентного увеличения объединенных расходов на телевизионную рекламу. Обратите внимание на наличие выброса в правой части гистограммы (компания Regal Communications — 2353,7%), который затмевает подробности, связанные с большинством других рекламодателей, и, по сути, сводит почти все компании в один столбик — от 0 до 500%

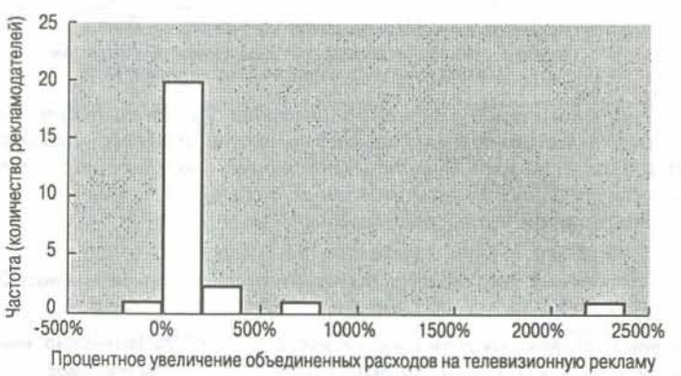


Рис. 3.6.2. Другая гистограмма для данных о всех 25 крупнейших рекламодателях, но с более узкими столбиками. Выброс справа на гистограмме по-прежнему скрывает большинство данных, хотя теперь мы ясно видим, что значения лежат в основном в пределе от 0 до 100%

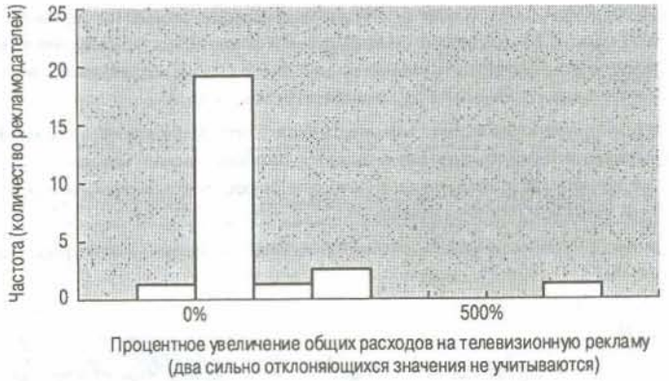


Рис. 3.6.3. Гистограмма изменений расходов 24 рекламодателей после исключения значения 2353,7% — наибольшего выброса (компания Regal Communications) и увеличения масштаба по горизонтали для получения более подробной картины. Теперь виден второй выброс — 684% (компания Himmel Group), который продолжает скрывать детали большей части набора данных

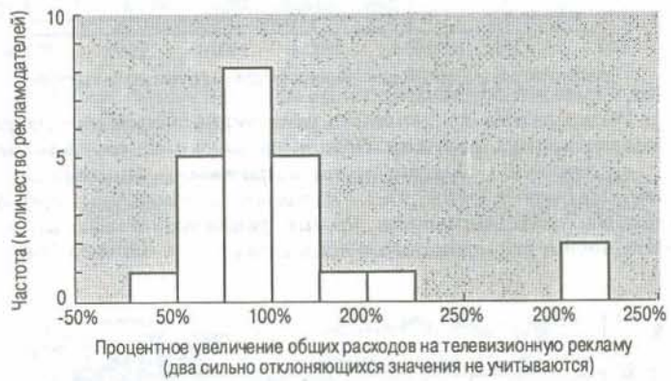


Рис. 3.6.4. Гистограмма изменений расходов 23 рекламодателей после исключения двух самых больших выбросов значений (компании Regal и Himmel). Теперь видны отдельные подробности, которые дают общую картину распределения изменений расходов (возможно, с двумя новыми выбросами справа)

Две компании имеют исключительно большое увеличение расходов, а именно Regal Communications (2 353,7%) и Himmel Group (684%).

Остальные компании дают типичное увеличение расходов от 0 до 75% (возможно, немного больше или меньше), за исключением двух компаний с высоким, около 200%, ростом расходов.

Данные этого анализа свидетельствуют о том, что затраты на рекламу сильно меняются каждый год. Крупные рекламодатели не имеют постоянной устойчивой стратегии, которая лишь немного корректируется каждый год. Большинство из 25 ведущих рекламодателей для телевидения, видимо, оказались в этом списке благодаря значительному увеличению своих расходов на рекламу по сравнению с предыдущим годом.

3.7. Гистограммы, построенные вручную: метод "ствол и листья"

В настоящее время наиболее эффективным является построение гистограмм с помощью компьютерных статистических пакетов программ. Однако иногда может возникнуть необходимость построить диаграмму вручную. Например, у вас нет возможности использовать компьютер до окончания проекта, а вам хочется проверить еще одну возможность путем изучения гистограммы. Кроме того, при небольшом количестве значений быстрее начертить гистограмму на бумаге, чем начинать вводить данные в компьютер. И наконец, когда вы самостоятельно чертите гистограмму, вы "приближаетесь к данным", интуитивно лучше чувствуете их, чем при вычислениях с помощью компьютера.

Наиболее простым способом построения диаграмм вручную является метод **"ствол и листья"**, при котором столбики гистограммы строятся путем записи одних чисел над другими (или рядом друг с другом). Преимущество этого метода построения состоит в том, что гистограмма растет прямо у вас на глазах и вы сразу видите, насколько информативны ваши данные.

Начните с определения того, какие начальные разряды записи значений необходимо включить в шкалу, чтобы отсечь таким образом мелкие подробности. Например, можно использовать для шкалы разряды миллионов и сотен тысяч, но не использовать десятки тысяч и более мелкие разряды. Затем добавьте к шкале *разряды следующего уровня* (десятки тысяч) и запишите каждое значение данных, строя колонку вверх (или в сторону), — это и будут столбики гистограммы.

Пример. *Служащие сферы общественного питания*

Рассмотрим данные о количестве служащих в фирмах общественного питания (из списка Fortune 1000), приведенные в табл. 3.7.1.

Используя при построении шкалы в качестве единицы измерения сто тысяч, следует начать с горизонтальной шкалы, со значениями из диапазона от 0 до 4.

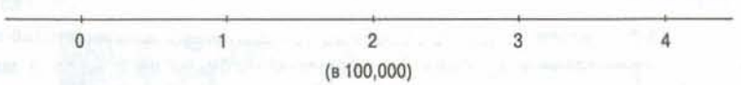
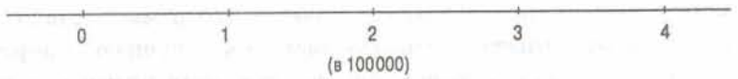


Таблица 3.7.1. Количество служащих в фирмах общественного питания

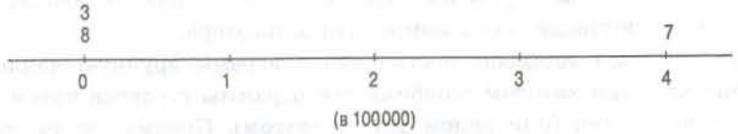
Фирма	Количество служащих	Фирма	Количество служащих
PepsiCo	471 000	Morrison Restaurants	33 000
McDonald's	183 000	Shoney's	30 000
Aramark	133 000	Family Restaurants	51 700
Flagstar	90 000	Foodmaker	26 170
Wendy's International	44 000	Brinker International	38 000

Данные из *Fortune*, 1995, May 15, p. F-52

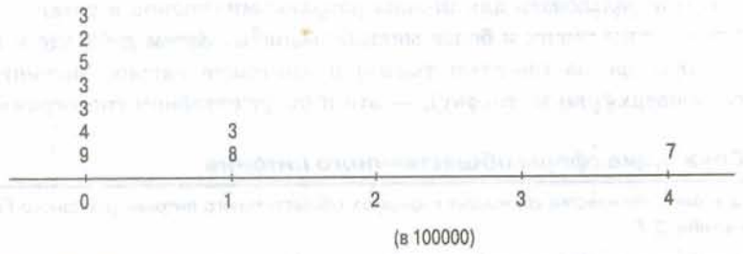
Как записать первое значение данных, представляющее количество служащих фирмы PepsiCo, равное 471 000? На первом месте в записи числа стоит цифра 4, выражающая сотни тысяч. Поэтому берем следующую цифру 7 и ставим ее на горизонтальной шкале над цифрой 4. Поскольку следующая цифра (1, обозначающая тысячи) менее важна, ее можно не наносить на шкалу. В результате получим следующее.



Следующее значение, 183000 служащих фирмы McDonald's, записывается как число 8 (десятки тысяч) над 1 (сотни тысяч). Далее, 133000 служащих фирмы Aramark записываем как цифру 3 над 1, размещая цифры одна над другой. После записи этих двух значений получаем:



Записывая таким образом новые значения, мы получаем колонки, которые растут вверх. Гистограмма растет у вас на глазах. Это более понятно (и более информативно), чем просто производить вычисления и подсчет частот до того, как будет получен какой-либо результат. Окончательная гистограмма, построенная методом "ствол и листья", будет выглядеть так:



Чтобы представить эту гистограмму в обычном виде, нужно просто заменить колонки цифр на столбики, как показано на рис. 3.7.1.

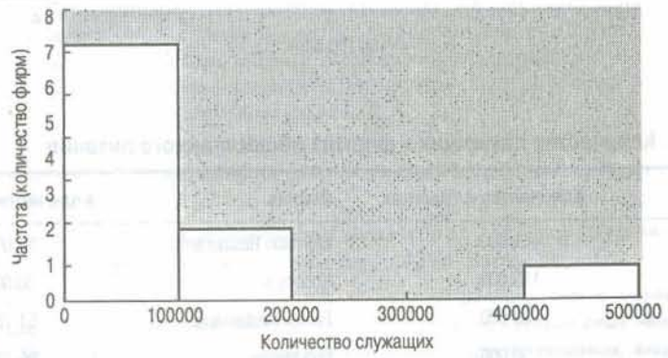


Рис. 3.7.1. Обычная гистограмма, в которой колонки чисел гистограммы "ствол и листья" заменены на столбики такой же высоты

Многие компьютерные программы позволяют строить такие гистограммы "ствол и листя" , у которых колонки чисел растут не вверх, а в сторону. Обратите внимание, что в компьютерных гистограммах "ствол и листья" каждая группа делится на две части: например, верхняя строка содержит значения от 10000 до 49999, а вторая строка — значения от 50000 до 99999 служащих.

Гистограмма типа "ствол и листья" для количества служащих, N=10.

Один лист = 10 000

5	0 23334
2	0 59
1	1 3
1	1 8
0	2
0	2
0	3
0	3
0	4
1	4 7

Обе гистограммы (и типа "ствол и листья", и обычная) демонстрируют основную информацию о наборе данных. Количество служащих крупных фирм общественного питания находится в диапазоне от нескольких десятков тысяч до почти полумиллиона. Небольшое количество данных не позволяет сделать точный вывод, но, похоже, что распределение достаточно асимметрично.

3.8. Дополнительный материал

Резюме

Набор данных простейшего вида представляет собой **список чисел**, содержащих некоторую информацию (единственная статистическая переменная), измеренную для каждого изучаемого объекта (каждой элементарной единицы). Такой список чисел может быть представлен либо действительно в виде списка, либо в виде таблицы, где записано, сколько раз каждое из значений повторяется в списке.

Первым шагом в анализе списка чисел является изучение гистограммы, которая дает представление об основных свойствах набора данных, таких как типичные значения, особые значения, концентрация, распределение значений, характер данных и наличие в данных отдельных групп значений. **Гистограмма** представляет частоты в виде столбиковой диаграммы, расположенной над числовой осью и показывающей, сколько раз различные значения встречаются в наборе данных. **Числовая ось** представляет собой прямую линию, обычно горизонтальную, с нанесенными под ней числами, образующими шкалу.

Нормальное распределение представляет собой теоретическую гладкую в форме колокола гистограмму, без случайных отклонений. Ей соответствует идеальный набор данных, в котором большинство значений сконцентрировано в средней части диапазона, а оставшиеся значения симметрично с затуханием частоты расположены по обе стороны от вершины колокола. Набор данных имеет

нормальное распределение, если форма его гистограммы близка к идеальной гладкой в форме колокола кривой, возможно, с некоторыми случайными отклонениями. Нормальное распределение играет важную роль в теории и практике статистического анализа.

Асимметричное (скошенное) распределение не является ни симметричным, ни нормальным, поскольку значения данных с одной стороны затухают более резко, чем с другой. Асимметричные распределения очень часто встречаются в бизнесе. К сожалению, большинство статистических методов не применимы к сильно скошенным распределениям.

Преобразование заключается в замене каждого значения другим числом (например, логарифмом этого значения) с целью упрощения статистического анализа. **Логарифмирование** часто преобразует асимметрию в симметрию, поскольку позволяет растянуть шкалу в окрестности нуля, растягивая по шкале все сгруппированные вместе малые значения. Логарифмирование также группирует большие значения, растянутые на правом конце исходной шкалы. Логарифмировать можно только положительные числа. Для правильной интерпретации результата логарифмирования необходимо учитывать, что равным расстояниям на логарифмической шкале соответствуют на исходной шкале равные процентные увеличения, а не равные увеличения значений (как, например, объем финансов в долларах).

Если на гистограмме четко видны две отдельные группы, то это говорит о **бимодальном распределении** данных. Важно уметь определять наличие бимодального распределения, чтобы предпринимать соответствующие действия при анализе. Возможно, выяснится, что вас интересует только одна из этих групп данных, а вторую можно не рассматривать. Возможно, придется вносить в анализ определенные изменения, чтобы справиться с этой более сложной ситуацией.

Иногда данные могут содержать **выбросы (сильно отклоняющиеся значения)**, т.е. одно или несколько таких значений, которые, по-видимому, не принадлежат данному распределению, поскольку либо слишком велики, либо слишком малы. Выбросы затрудняют статистический анализ, поэтому их следует идентифицировать и обработать специально. Если выброс представляет собой просто ошибку, то ее следует исправить и продолжить анализ. Если ошибки нет, а значение сильно отличается от остальных значений из набора данных, то этот выброс можно либо исключить, либо не исключить из анализа. Если вы убедите себя и других, что выброс не принадлежит изучаемой системе данных, его можно исключить. Если вы не можете обосновать исключение выброса, может потребоваться выполнить два анализа: с выбросом и без него. В любом случае в отчете вам необходимо четко написать о наличии выброса и предпринятых действиях.

Лучше всего строить гистограммы с помощью компьютерных пакетов программ статистического анализа. Однако иногда необходимо (и даже желательно) построить диаграмму вручную. В методе "ствол и листья" столбики гистограммы строят путем записи чисел одного над другим (или рядом с другим). Поскольку значения данных записаны в некоторой последовательности, у вас есть возможность делать интуитивные логические выводы о данных уже в процессе построения диаграммы.

Основные термины

- Последовательность чисел (list of numbers), 71
- Числовая ось (number line), 72
- Гистограмма (histogram), 73
- Нормальное распределение (normal distribution), 79
- Несимметричное (скошенное) распределение (skewed distribution), 81
- Преобразование (transformation), 85
- Логарифм (logarithm), 86
- Бимодальное распределение (bimodal distribution), 88
- Выброс (outlier), 91
- "Ствол и листья" (stem-and-leaf), 97

Контрольные вопросы

1. Что такое список чисел?
2. Назовите шесть свойств набора данных, которые можно увидеть на гистограмме.
3. Что такое числовая ось?
4. Чем отличается гистограмма от столбиковой диаграммы?
5. Что такое нормальное распределение?
6. Почему нормальное распределение играет важную роль в статистике?
7. Если реальный набор данных распределен нормально, то можно ли ожидать, что гистограмма будет иметь идеально гладкую форму в виде колокола? Обоснуйте свой ответ.
8. Все ли наборы данных подчиняются нормальному распределению?
9. Что такое асимметричное распределение?
10. В чем основная проблема асимметрии? Как эту проблему можно решить во многих случаях?
11. Как вы можете проинтерпретировать логарифм числа?
12. Что такое бимодальное распределение? Что следует предпринять в случае бимодального распределения?
13. Что такое выброс?
14. Почему важно описать в отчете, какие действия предпринимались в отношении выбросов?
15. Какие проблемы возникают при наличии выбросов значений?
16. В каких случаях выбросы можно не учитывать и анализировать только остальные данные?
17. Предположим, что в вашем наборе данных есть выбросы. Вы планируете проанализировать данные дважды: с выбросами и без них. Какой результат вас больше устроит? Почему?

18. Что такое гистограмма "ствол и листья"?

19. В чем заключаются преимущества гистограммы "ствол и листья"?

Задачи

1. В соответствии с программой контроля качества измерены значения электрического напряжения для исходных компонентов. Какую форму имеет распределение этих значений, представленных гистограммой на рис. 3.8.1?
2. Какую форму имеет распределение значений доли прибыли в цене потребительских товаров, представленное гистограммой на рис. 3.8.2?
3. Какую форму имеет распределение значений объемов продаж (в тысячах единиц) по регионам, представленное гистограммой на рис. 3.8.3?
4. Какую форму имеет распределение значений продолжительности пребывания в больнице (в днях), представленное гистограммой на рис. 3.8.4?
5. Рассмотрим гистограмму на рис. 3.8.5, которая показывает эффективность последних договоров на техническое обслуживание, представленную как норма прибыли.

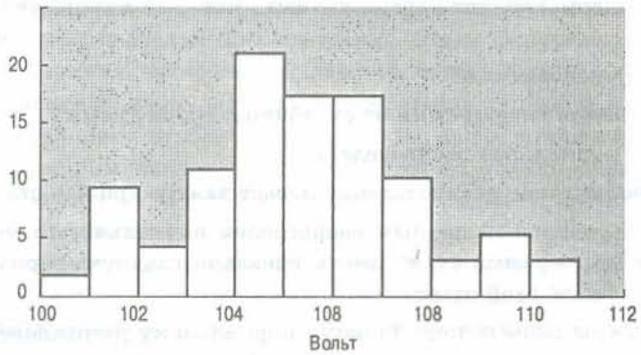


Рис. 3.8.1. Гистограмма значений напряжения для исходных компонентов

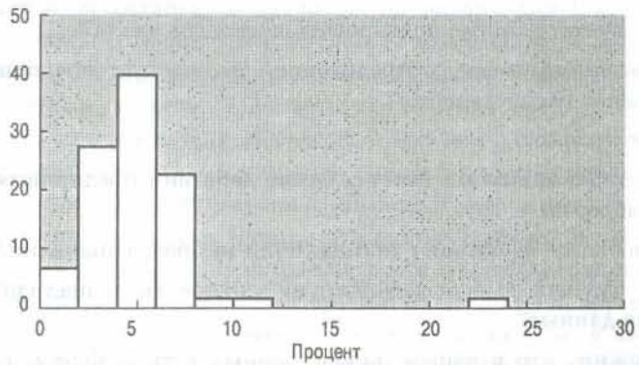


Рис. 3.8.2. Гистограмма распределения значений доли прибыли в цене потребительских товаров

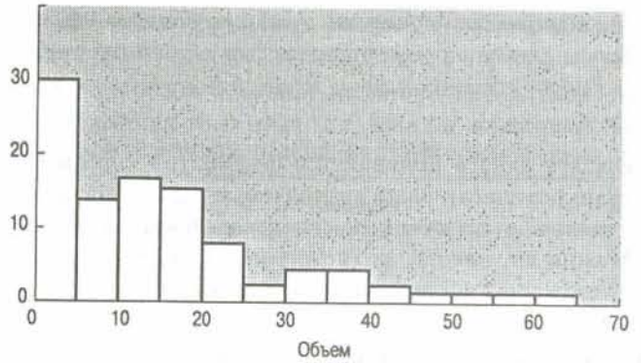


Рис. 3.8.3. Гистограмма распределения значений объемов продаж

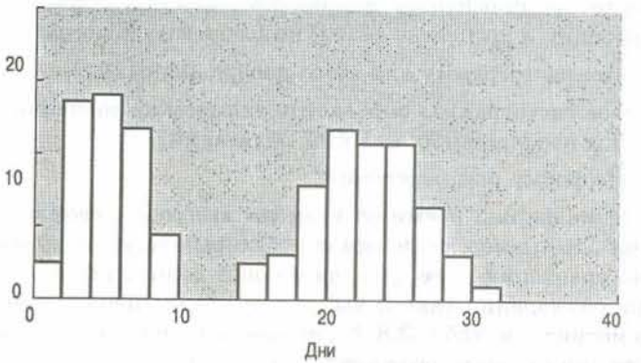


Рис. 3.8.4. Гистограмма продолжительности пребывания пациента в больнице

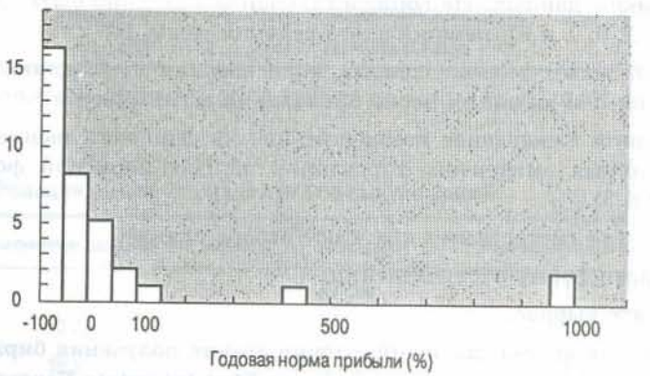


Рис. 3.8.5. Гистограмма распределения эффективности последних договоров на техническое обслуживание

9. Рассмотрим CREF (College Retirement Equities Fund), который управляет пенсионными счетами служащих небюджетных образовательных и научно-исследовательских организаций. CREF управляет большим и диверсифицированным портфелем акций, объемом приблизительно 67 миллиардов долларов. Инвестиции в магазины мебели и домашнего интерьера составляют 0,18% этого портфеля. Данные о рыночной стоимости этих инвестиций приведены в табл. 3.8.4.

   а) Постройте гистограмму набора данных.

   б) На основе гистограммы опишите распределение инвестиций CREF в магазины мебели и домашнего интерьера.

   в) Опишите форму распределения. В частности, укажите, это распределение симметричное или скошенное (асимметричное)?

   г) Вычислите логарифм каждого из значений данных.

Таблица 3.8.3. Активные эмитенты фондового рынка

Фирма	Изменение	Фирма	Изменение	Фирма	Изменение
Micron Tch	, 250	Citicorn	–3,625	ArcherDan	0,375
Compaq	, 750	SGS Thomson	2,375	EMC Cp	–0,375
IBM	– 0,500	Digital Eqp	2,500	FordMotor	–0,625
WalMart	, 000	Kmart	–0,250	GraceWR	–8,500
Motorola	, 375	TelefMex	0,250	FstInterste	34,250

Данные взяты из *The Wall Street Journal*, 1995, October, 19, p. C2.

Таблица 3.8.4. Инвестиции CREF

Магазин	Рыночная стоимость портфеля, тыс. дол.	Магазин	Рыночная стоимость портфеля, тыс. дол.
Australia Gas Light Co.	3463	Lechters, Inc.	293
Bed Bath & Beyond, Inc.	26 445	Linens N Things, Inc.	315
Best Buy, Inc.	1304	Maxim Group, Inc. (The)	706
Bombay, Inc.	1671	Microage, Inc.	52
Compucom Systems, Inc	71	Musicland Stores, Inc.	2843
CompUSA, Inc.	29 816	Pier 1 Imports, Inc.	29 530
Egghead. Com, Inc.	1007	Rex Stores Corp.	2521
Ethan Allen Interiors, Inc.	335	Sun Television & Appliances, Inc.	416
Good Guys, Inc.	2814	Sunbeam Corp.	5346
Heilig Meyers Co.	192	Tandy Corp.	67 305
Inacom Corp.	600	Trans World Entertainment Corp.	5593
JD Group Ltd.	398	Williams-Sonoma, Inc.	18 822

Данные взяты из CREF Semi-Annual Report, 1998, June, 30, p. 32.

   д) Постройте гистограмму для логарифмов значений.

   е) Опишите форму распределения логарифмов значений. В частности, укажите, это распределение симметричное или скошенное (асимметричное)?

10. Рассмотрим процентное изменение доходов компаний из списка Fortune 500, производящих фотоаппаратуру, научно-исследовательское и измерительное оборудование (табл. 3.8.5).

   а) Постройте гистограмму набора данных.

   б) Опишите форму распределения.

   в) Компания Varian Associates имеет наибольшее снижение дохода (отрицательное увеличение –10,9%) и на первый взгляд должна чем-то отличаться от других компаний. Исходя из построенной в п. "а" гистограммы, скажите, является ли фирма Varian выбросом? Обоснуйте свой вывод.

Таблица 3.8.5. Процентное изменение доходов компаний, производящих фотоаппаратуру, научно-исследовательское и измерительное оборудование (1997 г. по сравнению с 1996 г.)

Компания	Изменение доходов, %
Minnesota Mining & Mfg.	5,9
Eastman Kodak	–9,4
Honeywell	9,8
Baxter International	12,9
Thermo Electron	21,3
Becton Dickinson	1,5
Medtronic	12,4
Polaroid	–5,7
Tektronix	9,7
Bausch & Lomb	–0,6
Boston Scientific	28,1
EG&G	–1,8
Varian Associates	–10,9
Guidant	26,7
Perkin-Elmer	9,8
Teradyne	8,1
C. R. Bard	1,6
Beckman Instruments	16,5
United States Surgical	5,3

Данные взяты из *The Fortune 500 Industry List* по адресу `http://www.pathfinder.com/fortune/fortune500/ind24.html`, 11/12/1998.



11. Постройте гистограмму "ствол и листья" для набора данных из табл. 3.8.5.
12. Рассмотрите расходы на лечение заболеваний сердца в больницах района Puget Sound (табл. 3.8.6).
    а) Постройте гистограмму для этого набора данных.
    б) Опишите форму распределения.
13. Рассмотрим размеры вознаграждений, выплаченных главным должностным лицам фирм, производящих продукты питания (табл. 3.8.7).
    а) Постройте гистограмму для этого набора данных.
    б) Опишите форму распределения.
14. Радиостанции имеют свои стратегии работы, различаются своими программами, но всех их объединяет необходимость иметь аудиторию с целью привлечения рекламодателей. В табл. 3.8.8 приведены данные о проценте слушателей радиостанций в районе Seattle-Tacoma (в среднем возраст слушателей от 12 лет и старше, время вещания с 6 утра до полуночи всю неделю).
    а) Постройте гистограмму для набора данных.
    б) Опишите форму распределения.

Таблица 3.8.6. Расходы на лечение заболеваний сердца в больницах района Puget Sound (без учета оплаты врача)

Больница	Расходы, дол.	Больница	Расходы, дол.
Affiliated Health Services	6415	Overlake Medical Center	6364
Allenmore Community Hospital	5355	Providence General Medical Center	5235
Auburn Regional Medical Center	7189	Providence Saint Peter Hospital	5527
Cascade Valley Hospital	4690	Providence Seattle Medical Center	7222
Children's Hospital & Medical Center	8585	Puget Sound Hospital	9351
Columbia Capital Medical Center	6739	Saint Clare Hospital	6628
Community Memorial Hospital	4906	Saint Francis Community	6235
Evergreen Hospital Medical Center	5805	Saint Joseph Hospital	7110
Good Samaritan Hospital	4762	Saint Joseph Medical Center	6893
Group Health Central Hospital	3289	Stevens Memorial Hospital	5730
Group Health Eastside Hospital	2324	Swedish Medical Center	7661
Harborview Medical Center	7107	Tacoma General Hospital	5835
Harrison Memorial Hospital	5617	University of Washington Medical Center	7893
Highline Community Hospital	6269	Valley General Hospital	4279
Island Hospital	4811	Valley Medical Center	4863
Mary Bridge Children's Health Center	5582	Virginia Mason Medical Center	5773
Northwest Hospital	4759	Whidbey General Hospital	4142

Источник: *Book of Lists* 1998, Puget Sound Business Journal, Vol. 18, №33. Источник данных: отдел охраны здоровья штата Вашингтон.

Таблица 3.8.7. Вознаграждения руководителям, выплаченные фирмами, производящими продукты питания

Фирма	Вознаграждение, дол.	Фирма	Вознаграждение, дол.
Archer-Daniels-Midland	3 171 000	Kellogg	1 489 000
Campbell Soup	1 810 000	Pet	1 023 000
Conagra	1 600 000	Quaker Oats	1 398 000
CPC International	1 202 000	Ralston Purina Group	1 363 000
General Mills	850 000	Sara Lee	1 736 000
Heinz	895 000	Sysco	1 015 000
Hershey Foods	897 000	Tyson Foods	1 174 000
Hormel Foods	985 000	Wrigley	475 000

Данные взяты из: "Executive Compensation Scoreboard", *Business Week*, 1995, April, 24, p. 102.

Таблица 3.8.8. Распределение рынка между радиостанциями Сиэтла

Радиостанции	Программа (тип музыки)	Процент слушателей в возрасте от 12 лет и старше
KIXI - AM	Хиты 50–60-х годов	4,5
KBSG - FM - AM	Хиты 60–70-х годов	5,5
KJR - FM	Хиты 70-х годов	3,8
KLSY - FM	Современная для взрослых	4,2
KLSZ - FM	Современная для взрослых	4,0
KRWM - FM	Современная для взрослых	3,1
KMTT - FM - AM	Альтернативная для взрослых	3,5
KRWX - AM	Новости	1,7
KCMS - FM	Христианская (духовная) музыка	1,6
KCIS - FM	Христианская (духовная) музыка	0,4
KZOK - FM	Классический рок	5,4
KING - FM	Классическая	3,7
KMPS - FM - AM	Кантри	5,0
KRPM - FM - AM	Кантри	3,2
KYCW - FM	Кантри	3,2
KWJZ - FM	Современный джаз	2,7
KIRO - AM	Ток-новости	6,3
KOMO - AM	Музыкальные ток-новости	2,6
KISW - FM	Рок	4,0
KNDD - FM	Рок	4,6
KJR - AM	Ток-спорт	1,5
KIRO - FM	Ток-новости	2,3

Окончание табл. 3.8.8

Радиостанции	Программа (тип музыки)	Процент слушателей в возрасте от 12 лет и старше
KVI - AM	Ток-новости	4,9
KUBE - FM	Топ 40 / Ритм	6,0

Данные взяты из *The Seattle Times*, 1995, October, 20, p. F3. Источник: Arbitron Co (авторские права защищены).

15. В табл. 3.8.9 приведены доходы некоторых фирм за 1995 год.
    а) Постройте гистограмму для этого набора данных.
    б) Опишите форму распределения.
16. Многие люди не представляют, сколько стоят ритуальные услуги и насколько их стоимость различается в разных фирмах. Рассмотрим представленные в табл. 3.8.10 цены на оказание ритуальных услуг (исключая затраты на гроб и рытье могилы) для района Puget Sound штата Вашингтон.
    а) Постройте гистограмму для этого набора данных.
    б) Опишите форму распределения.
17. Когда в 1986 году был пересмотрен налоговый кодекс IRS, Конгресс освободил от налога некоторые корпорации. В табл. 3.8.11 содержатся данные о потерях поступлений в бюджет США в результате применения этих переходных правил для ряда корпораций.
    а) Постройте гистограмму для этого набора данных.
    б) Опишите форму распределения.

Таблица 3.8.9. Доходы некоторых фирм

Фирма	Доходы, млн дол.	Фирма	Доходы, млн дол.
AST Research	403	Alouette Cosmetics	3
Aasche Transport Svcs	12	American RE	39
Access Health	11	American Water Works	223
Accustaff	71	AMVESTORS Financial	4
ACMAT	11	ANCOR Communications	1
Actrade International	5	AON	100
ADAC Laboratories	50	APRIA Healthcare Group	278
Advanced Logic Research	53	ARCTCO	166
Advantage Bancorp	2	ASANTE Technologies	17
Air Canada	1315	Bally Entertainment	287
Ajay Sports	3	Bankers Life Holdings	31
AKZO Nobel	5219	Battery Technologies	6
Align-Rite International	8	Battle Mountain Gold	65
ALLMERICA & Cas	43	Belmont Homes	37

Данные взяты из первых двух колонок Digest of Earnings Reports, *The Wall Street Journal*, 1995, November, 3, p. C6.

Таблица 3.8.10. Стоимость обычных ритуальных услуг

Похоронное бюро	Стоимость, дол.
Bleitz	2180
Bonney-Watson	2250
Butterworth's Arthur A. Wright	2265
Dayspring & Fitch	1795
Evergreen-Washelli	1895
Faull-Stockes	2660
Flintoft's	2280
Green	3195
Price-Helton	2995
Purdy & Walters at Floral Hills	2665
Southwest Mortuary	2360
Yahn & Son	2210

Источник: *The Seattle Times*, 1996, December, 11, p. D5.

Таблица 3.8.11. Освобождение от некоторых налогов после пересмотра в 1986 году Налогового кодекса

Фирма	Снижение бюджетных поступлений, млн дол.	Фирма	Снижение бюджетных поступлений, млн дол.
Paramount Cards	7	New England Patriots	6
Banks of Iowa	7	Ireton Coal	18
Ideal Basic Industries	0	Ala-Tenn Resources	0
Goldrus Drilling	13	Metropolitan-First Minnesota Merger	9
Original Appalachian Artworks	6	Texas Air/Eastern Merger	47
Candle Corp.	13	Brunswick	61
S. A. Horvitz Testamentary Trust	1	Liberty Bell Park	5
Green Bay Packaging	2	Beneficial Corp.	67

Данные взяты из "Special Exemptions in the Tax Bill, as Disclosed by the Senate", *New York Times*, 1986, September, 27, p. 33. Указанные фирмы сгруппированы под заголовком "Переходные правила для общих резервов корпораций".

18. Продолжите работу со снижением бюджетных поступлений (табл. 3.8.11).
    а) Найдите логарифм каждого значения из набора данных. Пропустите две фирмы с нулевым снижением бюджетных поступлений.
    б) Постройте гистограмму для этого нового набора данных.
    в) Опишите форму распределения.

г) Сравните результаты этого анализа преобразованных значений данных с результатами анализа исходных данных (задача 17).

19. Ниже приведено количество электромоторов, отбракованных из-за низкого качества, в каждой из выпущенных в последнее время партий (размер партии составляет 250 единиц).

3, 2, 7, 5, 1, 3, 1, 7, 0, 6, 2, 3, 4, 1, 2, 25, 2, 4, 5, 0, 5, 3, 5, 3, 1, 2, 3, 1, 3, 0, 1, 6, 3, 5, 41, 1, 0, 6, 4, 1, 3.

а) Постройте гистограмму для этого набора данных.

б) Опишите форму распределения.

в) Определите выброс(ы) значений.

г) Исключите выброс(ы) и постройте гистограмму для оставшихся значений.

д) Сделайте вывод о качестве продукции фирмы.

20. Рассмотрите стоимость недельного проката автомобиля (с ручной коробкой передач) с небольшим размером освобождения от оплаты убытков в случае аварии (табл. 3.8.12).

а) Постройте гистограмму для этого набора данных.

б) Опишите форму распределения.

21. Постройте гистограмму процентных ставок, предложенных банками на депозитные сертификаты, и опишите форму распределения:

9,9%; 9,5%; 10,3%; 9,3%; 10,4%; 10,7%; 9,1%; 10,0%; 8,8%; 9,7%; 9,9%; 10,3%; 9,8%; 9,1%; 9,8%

22. Постройте гистограмму рыночной стоимости фондов ваших главных конкурентов (в миллионах долларов) и опишите форму распределения:

3,7; 28,3; 10,6; 0,1; 9,8; 6,2; 19,7; 23,8; 17,8; 7,8; 10,8; 10,9; 5,1; 4,1;2,0; 24,2;9,0; 3,1; 1,6; 3,7; 27,0; 1,2; 45,1; 20,4; 2,3

23. Рассмотрите размер заработной платы (в тысячах долларов) группы менеджеров:

177, 54, 98, 57, 209, 56, 45, 98, 58, 90, 116, 42, 142, 152, 85, 53, 52, 85, 72, 45, 168, 47, 93, 49, 79, 145, 149, 60, 58

а) Постройте гистограмму для этого набора данных.

Таблица 3.8.12. Стоимость проката автомобиля

Страна	Стоимость проката, дол.	Страна	Стоимость проката, дол.
Австрия	239	Нидерланды	194
Англия	229	Норвегия	241
Бельгия	179	Испания	154
Дания	181	Швеция	280
Франция	237	Швейцария	254
Ирландия	216	Германия	192
Италия	236		

б) Опишите форму распределения.

в) Исходя из гистограммы, укажите, какие значения размера заработной платы типичны для этой группы?

24. Рассмотрите размер последних заказов потребителей (в тысячах долларов):

31, 14, 10, 3, 17, 5, 1, 17, 1, 2, 7, 12, 28, 4, 4, 10, 4, 3, 9, 28, 4, 3

а) Постройте гистограмму для этого набора данных.

б) Опишите форму распределения.

25. Начертите гистограмму для цен (в долларах) одной пачки конвертов в различных магазинах и опишите форму распределения:

4,40; 4,20; 4,55; 4,45; 4,40; 4,10; 4,10; 3,80; 3,80; 4,30; 4,90; 4,20; 4,05

26. Рассмотрите следующий список размеров части рынка, которую занимает ваше изделие в 20 главных регионах:

0,7%; 20,8%; 2,3%; 7,7%; 5,6%; 4,2%; 0,8%; 8,4%; 5,2%; 17,2%; 2,7%; 1,4%; 1,7%; 26,7%; 4,6%; 15,6%; 2,8%; 21,6%; 13,3%; 0,5%

а) Постройте подходящую гистограмму для данного набора данных.

б) Опишите форму распределения.

27. Рассмотрите процентное изменение курса доллара по отношению к другим иностранным валютам в течение четырех недель (табл. 3.8.13).

а) Постройте гистограмму для данного набора данных.

б) Опишите форму распределения.

28. Рассмотрите следующий перечень цен (в долларах) набора из двенадцати таблеток (по 60 мг) лекарства "Тайленол № 4" с кодеином[17], который отпускают по рецепту в различных аптеках:

6,75; 12,19; 9,09; 9,09; 13,09; 13,45; 7,89; 12,00; 10,49; 15,30; 13,29.

а) Постройте гистограмму для этих цен.

б) Опишите форму распределения.

в) Прокомментируйте следующее утверждение: "Не имеет особого значения, где вы покупаете лекарство по рецепту".

29. Используя данные из задачи 24 главы 2 о транспортном индексе Dow Jones Transportation Average, выполните следующее:

а) Постройте диаграмму "ствол и листья" для процентного изменения по сравнению с 31 августа 1998 года.

б) Постройте гистограмму для процентного изменения по сравнению с 31 августа 1998 года.

в) Опишите форму распределения.

30. Используя данные из задачи 25 главы 2 о транспортном индексе Dow Jones Transportation Average, выполните следующее:

а) Постройте диаграмму "ствол и листья" для чистого изменения.

б) Постройте гистограмму для чистого изменения.

[17] Данные взяты из Gilje S. "What Health-Care Revision Means to Prescription Drug Sales", *The Seattle Times*, 1993, February 28, p. K1 (набор данных создан Morningstar C. и Hendrickson M.).

Таблица 3.8.13. Процентное изменение курса доллара

Иностранная валюта	Изменение курса доллара, %	Иностранная валюта	Изменение курса доллара, %
Англия	–3,7	Сингапур	–1,5
Бельгия	–5,3	Франция	–4,9
Япония	–6,7	Южная Корея	–1,0
Бразилия	26,0	Гонконг	0,0
Мексика	–1,2	Тайвань	–0,1
Нидерланды	–5,1	Италия	–4,7
Канада	–1,9	Германия	–5,1

в) Опишите форму распределения.

г) Постройте диаграмму "ствол и листья" для процентного изменения.

д) Постройте гистограмму для процентного изменения.

е) Опишите форму распределения.

Упражнения с использованием базы данных

Обратимся к базе данных, приведенной в приложении А.

1. Для заработной платы служащих:

   а) Постройте гистограмму.

   б) Опишите форму распределения.

   в) Обобщите информацию о распределении, указав также размеры наименьшей и наибольшей заработной платы.

2. Для возраста служащих:

   а) Постройте гистограмму.

   б) Опишите форму распределения.

   в) Обобщите информацию о распределении.

3. Для стажа работы служащих:

   а) Постройте гистограмму.

   б) Опишите форму распределения.

   в) Обобщите информацию о распределении.

4. Для заработной платы служащих разного пола:

   а) Постройте гистограмму только для мужчин.

   б) Постройте гистограмму для женщин, используя тот же масштаб, что и в п. "а", с целью сравнения заработной платы мужчин и женщин.

   в) Сравните эти два распределения заработной платы и напишите резюме, указав наблюдаемые на гистограммах различия заработной платы мужчин и женщин[18].

[18] Подобные статистические методы сравнения двух групп будут представлены в главе 10.

Проекты



Постройте гистограмму для каждого из трех наборов данных, имеющих отношение к вашим интересам в бизнесе. Подберите интересующие вас данные в Internet, *The Wall Street Journal* или в отчетах вашей фирмы. Каждый набор данных должен содержать не меньше 15 чисел. Для каждого набора данных напишите страничку комментария (включая гистограмму), указав следующее:

а) Какова форма распределения?

б) Есть ли какие-либо выбросы значений? Что нужно сделать, если они есть?

в) Обобщите информацию о распределении.

г) Что вы узнали, изучив гистограмму?

Ситуация для анализа

Необходимость контроля производственных потерь

"Этот Оуэн выбрасывает наши деньги на ветер! — громко заявил Биллингс на совещании. — У меня есть доказательства. Вот гистограмма стоимости использованного сырья. Четко видны две группы, причем Оуэн тратит на сырье на несколько сотен долларов больше, чем Парсел".

Вы ведете совещание, и оно проходит более эмоционально, чем хотелось бы. Чтобы ввести собрание в более спокойное русло, вы вежливо пытаетесь смягчить обсуждение и досконально обдумать решение. Вы не одиноки в своем желании. Есть предложение изучить данный вопрос и внести его в повестку дня следующего совещания.

Вы знаете, как, впрочем, и большинство других, что Оуэн имеет репутацию беспечного человека. Однако вы никогда не ставили этот порок на первое место, и вам хотелось бы отложить оценку Оуэна как раз потому, что другие завистливо подбрасывают такое предложение, и потому, что Оуэна уважают за компетентность и трудолюбие. Вам также известно, что Биллингс и Парсел — хорошие приятели. В этом, конечно, нет ничего плохого, но все же лучше познакомиться со всей доступной информацией перед тем, как делать окончательный вывод.

После совещания вы просите Биллингса прислать вам по электронной почте копию данных. Но он присылает вам только первые две колонки (затраты на материалы), которые вы видите ниже, и они вам уже знакомы. В вашем компьютере уже есть отчет, включающий все три колонки, приведенные ниже. Теперь вы готовы потратить время на подготовку совещания, чтобы провести его на следующей неделе.

Стоимость сырья, дол.	Ответственный менеджер	Стоимость продукции, дол.	Стоимость сырья, дол.	Ответственный менеджер	Стоимость продукции, дол.
1459	Оуэн	4669	1434	Оуэн	4589
1502	Оуэн	4806	1127	Парсел	3606

Окончание таблицы

Стоимость сырья, дол.	Ответственный менеджер	Стоимость продукции, дол.	Стоимость сырья, дол.	Ответственный менеджер	Стоимость продукции, дол.
1492	Оуэн	4774	1457	Оуэн	4662
1120	Парсел	3558	1109	Парсел	3549
1483	Оуэн	4746	1236	Парсел	3955
1136	Парсел	3635	1188	Парсел	3802
1123	Парсел	3594	1512	Оуэн	4838
1542	Оуэн	4934	1131	Парсел	3619
1484	Оуэн	4749	1108	Парсел	3546
1379	Оуэн	4413	1135	Парсел	3632
1406	Оуэн	4499	1416	Оуэн	4531
1487	Оуэн	4758	1170	Парсел	3744
1138	Парсел	3642	1417	Оуэн	4534
1529	Оуэн	4893	1381	Оуэн	4419
1142	Парсел	3654	1248	Парсел	3994
1127	Парсел	3606	1171	Парсел	3747
1457	Оуэн	4662	1471	Оуэн	4707
1379	Оуэн	4733	1142	Парсел	3654
1407	Оуэн	4502	1161	Парсел	3715
1105	Парсел	3536	1135	Парсел	3632
1126	Парсел	3603	1500	Оуэн	4800

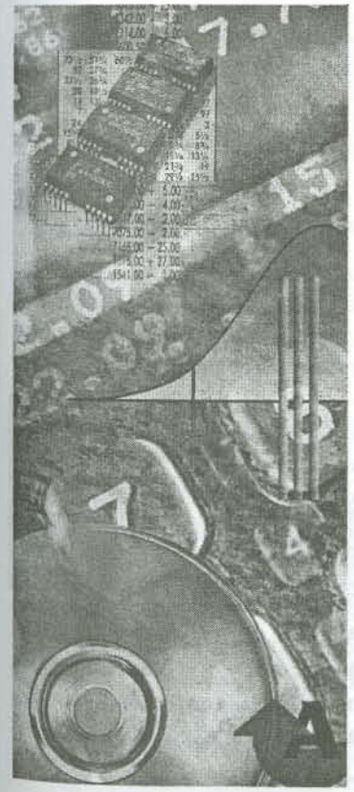
Вопросы для обсуждения

1. Является ли распределение стоимости сырья действительно бимодальным? Или эти данные можно рассматривать как одну нормально распределенную группу значений?
2. Согласуются ли гистограммы, построенные для Оуэна и Парсела отдельно, с утверждением Биллингса о том, что Оуэн тратит больше?
3. Нужно ли согласиться с Биллингсом на следующем совещании? Обоснуйте ваш ответ с помощью тщательного анализа имеющихся данных.

ГЛАВА 4

Обобщающие показатели: интерпретация типических значений и перцентилей

В сложных ситуациях один из самых эффективных способов "увидеть всю картину" заключается в **обобщении**, т.е. использовании одного или нескольких отобранных или рассчитанных значений для характеристики набора данных. Подробное изучение каждого отдельного случая само по себе не является статистической деятельностью[1], но обнаружение и идентификация особенностей, которые в целом характерны для рассматриваемых случаев, представляют собой статистическую деятельность, так как вся информация при этом рассматривается в целом.



Одна из целей статистики состоит в том, чтобы свести набор данных к одному числу (или двум, или нескольким), которое выражает наиболее фундаментальные свойства данных. Методы, наиболее подходящие для анализа одного списка чисел (т.е. одномерного набора данных), включают определение следующих показателей.

Среднее, *медиана* и *мода* — это различные способы выбора единственного числа, которое лучше всего описывает все числа в наборе данных. Такой представленный одним числом показатель называется *типическим значением*, или *центром* (также используют термин *мера центральной тенденции. — Прим. ред.)*.

Перцентиль (также используют термин *процентиль. — Прим. ред.*) обобщает информацию о *рангах*, характеризуя значение, достигаемое заданным процентом общего количества данных, после того, как данные упорядочиваются (ранжируются) по возрастанию.

[1] Если есть время для изучения каждого значения, может быть, это стоит сделать!

Стандартное отклонение — характеристика различий между значениями в наборе данных. Это понятие также называют *разбросом*, или *изменчивостью* (подробно об этом — в главе 5).

Как быть, если набор данных содержит отдельные значения, которые неадекватно описываются этими показателями? Такие выбросы (сильно отклоняющиеся значения) можно просто описать отдельно. Таким образом, можно охарактеризовать большой набор данных, обобщив основные свойства большинства его элементов и затем создав список исключений. Это позволяет достичь статистической цели эффективного описания большого набора данных с учетом особой природы отдельных элементов.

4.1. Чему равно наиболее типическое значение?

Простейшее обобщение любого набора данных представляет собой единственное число, которое наилучшим образом представляет все значения данных. Такое число можно было бы назвать *типическим значением* для данного набора данных. Если не все значения в наборе данных одинаковы, то мнения о "наиболее типическом" могут быть разными. Существуют три вида такой обобщающей меры.

1. *Среднее*, которое можно вычислять только для имеющих содержательный смысл чисел (для количественных данных).
2. *Медиана*, или серединная точка, которую можно вычислять как для упорядоченных категорий (порядковые данные), так и для чисел.
3. *Мода*, или наиболее часто встречающаяся категория, которую можно вычислять для неупорядоченных категорий (для номинальных данных), для упорядоченных категорий и для чисел.

Среднее: типическое значение для количественных данных

Среднее чаще всего используют как типическое значение списка чисел и вычисляют путем сложения всех чисел списка и деления полученной суммы на количество чисел в списке (количество элементарных единиц). Формула вычисления выборочного среднего (т.е. среднего выборки данных) имеет следующий вид:

Выборочное среднее

$$\text{Выборочное среднее} = \frac{\text{Сумма значений элементов данных}}{\text{Количество элементов данных}},$$

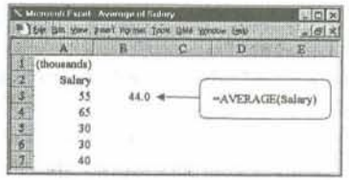
$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

где n — общее число элементов в списке данных, $X_1, X_2, \ldots X_n$ — непосредственно сами значения данных. Греческая прописная буква *сигма*, $\sum$, указывает на необходимость сложить все значения, которые записаны за ней, заменяя при этом индекс i значениями от 1 до n. Символ для записи среднего $\bar{X}$ произносится как "икс с чертой".

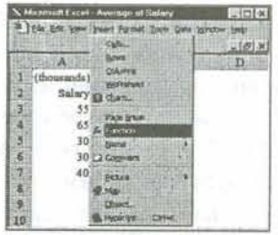
Например, среднее набора данных из трех чисел, 4, 9, 8, равно

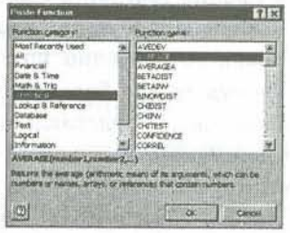
$$\frac{4+9+8}{3} = \frac{21}{3} = 7 .$$

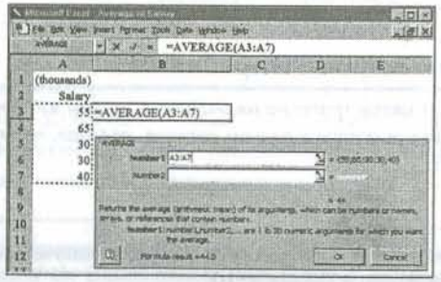
Ниже показано, как можно использовать функцию Average (СРЗНАЧ) электронных таблиц Excel для вычисления среднего списка чисел, расположенных в колонке непосредственно под заголовком Salary (Зарплата).



Система меню Excel поможет вам вычислить среднее (или значение другой статистической функции). Начните с выбора ячейки, куда вы хотите поместить среднее. Далее выберите в главном меню Insert⇨Function (Вставка⇨Функция), затем в качестве категории функции выберите Statistical (Статистические) и в качестве названия функции укажите Average (СРЗНАЧ). Появится диалоговое окно. Перетаскивая курсор мыши, выделите необходимый список чисел, а затем нажмите клавишу <Enter> для завершения процесса. Вот как это выглядит.







Понятие среднего не зависит от того, представляет ваш список чисел всю генеральную совокупность или же репрезентативную выборку из большей совокупности[2]. В то же время обозначения несколько различаются. Для всей генеральной совокупности количество элементов обозначают буквой N, а среднее — буквой μ (греческая буква "мю"). Процесс вычисления среднего одинаков как для генеральной совокупности, так и для выборки.

Поскольку при вычислении среднего значения данные суммируют, ясно, что среднее нельзя вычислять для качественных данных (нельзя складывать цвета или рейтинги долговых обязательств).

Среднее можно интерпретировать как равномерное распределение суммы всех значений между элементарными единицами. Таким образом, если каждое значение данных заменить средним, то общая сумма не изменится. Например, из базы данных служащих можно вычислить среднюю заработную плату служащих в Хьюстоне. Это среднее можно интерпретировать таким образом: если бы мы выплачивали всем служащим Хьюстона одинаковую заработную плату, не изменяя при этом общий фонд заработной платы, то значение этой заработной платы было бы равно среднему. Обратите внимание, что не следует рассматривать структуру уровня заработной платы, которая получена исходя из среднего, в качестве индикатора типичной заработной платы (особенно, когда вы имеете дело с фондом заработной платы как части бюджета).

Поскольку среднее сохраняет неизменной сумму при равномерном распределении значений, оно наиболее полезно в качестве обобщающего показателя при отсутствии экстремальных значений (выбросов), когда набор данных представляет собой более-менее однородную группу с элементами случайности. Если один служащий зарабатывает намного больше других, то среднее нельзя использовать в качестве обобщающего показателя. Хотя среднее и сохраняет неизменной общую сумму заработной платы, оно не будет хорошим показателем величины заработной платы отдельных служащих, так как среднее будет слишком высоким для большинства служащих и слишком низким для этого высокооплачиваемого работника.

Среднее является только обобщающей характеристикой, которая сохраняет общую сумму. Это свойство среднего особенно полезно в тех ситуациях, когда необходимо планировать общую сумму для большой группы. Сначала вычисляют среднее для меньшей выборки данных, представляющей большую группу. Затем полученное среднее можно умножить на количество отдельных элементов в этой большей группе. В результате получают оценку или прогноз суммы для большей по размеру совокупности. В целом, если необходимо определить общую сумму, можно использовать среднее.

Пример. *Сколько денег потратят потребители?*

Фирма интересуется, сколько в целом тратят на медицинские товары жители Кливленда. Анализ случайной выборки из трехсот человек, живущих в данном регионе, показал, что в прошлом месяце каждый из них потратил в среднем $6,58.

[2] Понятие выборки из генеральной совокупности является ключевым для статистического вывода и будет детально рассмотрено в главах 8–10.

Естественно, кто-то потратил больше, а кто-то меньше этого среднего количества денег. Вместо того чтобы работать со всеми 300 числами, мы используем среднее, чтобы определить типическое значение индивидуальных расходов каждого потребителя. Что особенно важно, умножив среднее значение расходов на численность населения Кливленда, мы получили приемлемую оценку суммарных расходов на медицинские товары жителей всего города[3]:

Оценка затрат на медицинские товары жителей Кливленда = (среднее значение расходов одного человека из выборки) (численность населения Кливленда) = ($6,58) (503 000) = $3 309 740.

Этот прогноз суммарных продаж, равный $3300000, является приемлемым и, вероятно, полезным. Однако это значение не является точным (в том смысле, что оно не отражает точную сумму потраченных денег). Далее, при изучении доверительных интервалов (в главе 9), вы узнаете, как учитывать статистическую ошибку, возникающую при распространении результата, полученного для выборки из 300 человек, на все население, состоящее из 503 000 человек.

Пример. *Сколько имеется бракованных деталей?*

Каждая партия изделий компании Globular Ball Bearing Company содержит 1000 изделий. Для проведения контроля качества изделий из произведенных за день 253 партий была взята случайным образом выборка, включающая 10 партий. Число бракованных изделий в каждой партии составило:

$$3, 8, 2, 5, 0, 7, 14, 7, 4, 1.$$

Среднее для этого набора данных:

$$\frac{3+8+2+5+0+7+14+7+4+1}{10}=\frac{51}{10}=5,1$$

демонстрирует, что в среднем каждая партия содержит 5,1 бракованных изделий. Иными словами, уровень брака составляет 5,1 изделия на 1000, или 0,51% (примерно полпроцента). Если распространить полученное среднее на все выпущенные за день 253 партии, то можно ожидать

$$5,1 \times 253 = 1290,3$$

бракованных изделий в дневном выпуске продукции, который составляет 253 000 изделий.

Чтобы показать, насколько среднее действительно является приемлемой обобщающей характеристикой списка чисел, на рис. 4.1.1 приведена гистограмма для этого набора данных из 10 чисел с обозначенным средним. Обратите внимание, насколько хорошо в середине данных расположено среднее, оно достаточно близко ко всем значениям данных.

Взвешенное среднее: учет важности

Взвешенное среднее (используют также термин **средневзвешенное**. — *Прим. ред.*) похоже на среднее, но позволяет присвоить различную важность (значимость), или "вес", каждому элементу данных. Взвешенное среднее дает возможность гибко определять систему важности отдельных элементов данных в том случае, когда их нельзя рассматривать как равноценные.

Если у фирмы три завода, при анализе пенсионных расходов не нужно брать простое среднее типических размеров пенсионных расходов на каждом из трех заводов как типическое значение общих пенсионных расходов, особенно, если заводы отличаются по размеру. Если численность служащих на одном заводе в два раза превышает численность служащих на другом заводе, по-видимому, бу-

[3] Здесь используется оценка численности населения в 1992 году по данным Бюро переписи населения США. *Statistical Abstract of the United States: 1994*, 114th ed. (Washington, D. C., 1994), p. 44.



Рис. 4.1.1. Гистограмма количества бракованных деталей в каждой из 10 партий (одна партия содержит 1000 изделий) с указанным средним значением (5,1)

дет разумным при вычислении обобщающего показателя учесть пенсионный фонд первого завода дважды. Средневзвешенное позволит вам обобщить данные, используя веса, определенные в соответствии с размером каждого завода.

Веса обычно представляют собой положительные числа, сумма которых равна 1. Не волнуйтесь, если первоначально вычисленная сумма весов не равна 1. Вы всегда сможете откорректировать значения весов, разделив каждый вес на сумму всех других весов. Исходные веса можно было бы определить исходя из численности служащих, рыночной стоимости или любого другого объективного показателя, а также можно воспользоваться субъективным методом (руководствуясь чьим-то личным мнением или мнением эксперта). Иногда легче выбирать веса, не заботясь, чтобы их сумма была равна 1, а затем преобразовать их, разделив каждый на общую сумму.

Предположим, вы решили вычислить средневзвешенное пенсионных расходов для трех заводов, присвоив веса в соответствии с численностью служащих. Если численность служащих равна 182, 386 и 697, то веса соответственно равны:

$$182/1\,265 = 0{,}144;$$
$$386/1\,265 = 0{,}305;$$
$$697/1\,265 = 0{,}551.$$

Обратите внимание, что значение веса получено путем деления численности служащих на данном заводе на общее количество служащих трех заводов — 182 + 386 + 697 = 1265. Сумма полученных весов, как это и требуется, равна 1[4]: 0,144 + 0,305 + 0,551 = 1.

Для вычисления взвешенного среднего каждый элемент данных умножают на присвоенный ему вес и суммируют полученные значения. Соответствующая формула имеет такой вид.

[4] Реально сумма может быть равна 0,999 или 1,001 (в зависимости от ошибок округления). Не обращайте на это внимание.

Взвешенное среднее

Взвешенное среднее = Сумма (вес умноженный на значение элемента) =

$$= \omega_1 X_1 + \omega_2 X_2 + \ldots + \omega_n X_n = \sum_{i=1}^{n} \omega_i X_i$$

где $\omega_1, \omega_2, \ldots, \omega_n$ — соответствующие веса, сумма которых равна 1. Вы можете считать обычное (не взвешенное) среднее также средневзвешенным, в котором все элементы данных имеют одинаковый вес, равный $1/n$.

Средневзвешенное значений 63, 47 и 98 с весами, равными 0,144; 0,305 и 0,551, соответственно, равно:

$$(0{,}144 \times 63) + (0{,}305 \times 47) + (0{,}551 \times 98) =$$
$$= 9{,}072 + 14{,}335 + 53{,}998 = 77{,}405$$

Обратите внимание, что, как и следовало ожидать, средневзвешенное отличается от обычного (не взвешенного) среднего этих трех значений (53 + 47 + 98)/3 = 69,333. При вычислении средневзвешенного наибольшее значение имеет вес 0,551 (что больше, чем одна треть суммарного веса). Вот почему в нашем случае средневзвешенное больше, чем обычное не взвешенное среднее.

Средневзвешенное лучше всего интерпретировать как среднее, используемое в ситуациях, когда одни элементы более важны, чем другие. Более важные элементы вносят больший вклад в значение средневзвешенного.

Пример. *Ваш средний балл*

Средний балл (GPA — grade point average) ваших результатов обучения в университете вычисляется как взвешенное среднее. Это связано с тем, что некоторые курсы оцениваются большим количеством очков и, следовательно, являются более важными по сравнению с другими. Вполне разумно, если курсу, который оценивается в два раза больше, чем другой, присваивается вдвое больший вес, и средний балл это отражает.

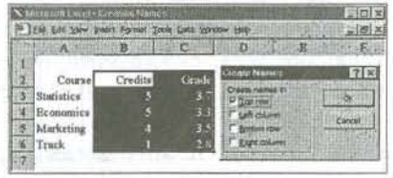
В разных университетах используют разные системы оценок. Предположим, что система оценок в вашем университете включает оценки от 0,0 (незачет) до 4,0 (отлично) и в конце семестра ваша карточка с оценками имеет такой вид.

Курс (Course)	Очки (Credits)	Оценка (Grade)
Статистика (Statistics)	5	3,7
Экономика (Economics)	5	3,3
Маркетинг (Marketing)	4	3,5
Спецкурс (Track)	1	2,8
Итого	15	

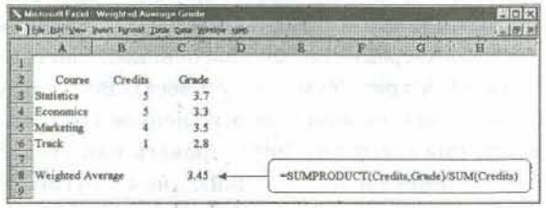
Веса можно вычислить, разделив количество очков по текущему курсу на 15 — общую сумму очков. Ваш средний балл рассчитывают как средневзвешенное ваших оценок, взвешенное в соответствии с количеством очков каждого из курсов:

$$\left(\frac{5}{15} \times 3{,}7\right) + \left(\frac{5}{15} \times 3{,}3\right) + \left(\frac{4}{15} \times 3{,}5\right) + \left(\frac{1}{15} \times 2{,}8\right) = 3{,}45.$$

Чтобы найти средневзвешенное с помощью Excel, вначале дайте названия каждой из колонок чисел. Проще всего это сделать, выделив все колонки, включая верхние заголовки, затем использовать команду Excel Insert⇨Name⇨Create (Вставка⇨Имя⇨Создать) и щелкнуть на кнопке OK. Это выглядит следующим образом.



Теперь используйте функцию Excel SUMPRODUCT (СУММПРОИЗВ), которая умножает количество очков курса на соответствующую оценку и складывает полученные результаты, а затем результат разделите на общее количество очков (чтобы сумма весов была равна 1). Вы получите средневзвешенное значение, равное 3,45.



К счастью, низкая оценка за спецкурс не сильно повлияла на ваш средний балл (GPA), равный 3,45, потому что вес этой оценки мал (всего 1 очко). Если бы эти четыре оценки были просто усреднены, то результат был бы ниже (3,33). Большая удача, что в решающий момент в конце семестра ваши оценки не слишком пострадали из-за вашего увлечения экономическими курсами!

Пример. *Стоимость капитала фирмы*

Стоимость капитала фирмы, понятие из области корпоративного финансирования, вычисляют как взвешенное среднее. Суть в том, что фирма увеличивает свои денежные средства посредством продажи различных ценных бумаг: акций, облигаций, векселей и т. д. Поскольку каждый вид ценной бумаги имеет свою собственную доходность (стоимость капитала), полезно объединить и обобщить различные уровни доходности в одно значение, представляющее собой совокупную стоимость капитала для этого набора ценных бумаг.

Стоимость капитала фирмы является простым средневзвешенным стоимости капитала по каждой ценной бумаге (доходность или процентная ставка), причем веса определяют в соответствии с полной рыночной стоимостью этих ценных бумаг. Например, если стоимость привилегированных акций составляет только 3% от рыночной стоимости выпущенных фирмой в обращение ценных бумаг, то ей следует присвоить низкий вес.

Рассмотрим ситуацию для Leveraged Industries, Inc., гипотетической фирмы со множеством долговых обязательств, образовавшихся в результате недавней деятельности по слиянию и приобретению[5].

[5] При вычислении стоимости капитала всегда используют текущее значение рыночной стоимости (а не стоимость покупки), поскольку рыночные значения показывают наиболее вероятную стоимость увеличиваемого фирмой капитала. Так, например, для облигаций следует использовать рыночную процентную ставку, а не купонную, основанную на номинальной стоимости, так как облигации могут не продаваться по номинальной стоимости. Кроме того, для получения общей рыночной стоимости выпущенных облигаций нужно умножить стоимость одной облигации на количество выпущенных облигаций. При изучении финансового курса очень важно научиться оценивать доход акций, который может быть получен на рынке.

Вид ценных бумаг	Рыночная стоимость, тыс. дол.	Доходность (норма прибыли), %
Обычная акция	100	18,5
Привилегированная акция	15	14,9
Облигации (ставка 9%)	225	11,2
Облигации (ставка 8,5%)	115	11,2
Итого	455	

Для каждого вида ценных бумаг разделите соответствующую рыночную стоимость на общую сумму, чтобы найти веса, которые дают пропорцию рыночной стоимости этого вида ценных бумаг[6].

Вид ценной бумаги	Вес
Обычная акция	0,220
Привилегированная акция	0,033
Облигации (9%)	0,495
Облигации (8,5%)	0,253

Для обычных акций вес равен 0,220, что в терминах рыночной стоимости означает, что на 22% фирма финансируется за счет обычных акций. Стоимость капитала можно вычислить, умножив значения рыночной доходности на веса и сложив полученные значения:

$$(0,220\times18,5)+(0,033\times14,9)+(0,495\times11,2)+(0,253\times11,2)=12,94$$

Стоимость (доходность) капитала Leveraged Industries, Inc. составляет 12,9%. Такое средневзвешенное объединяет значения стоимости (доходности) отдельных видов ценных бумаг (18,5%, 14,9% и 11,2%) в одно число.

Результат (12,9%) не изменился бы, если бы вы объединили два вида облигаций, рассмотрев их как один элемент с общей рыночной стоимостью $340 000 и доходностью 11,2% (что это так, можно проверить, произведя расчеты). Это связано с тем, что отличие между облигациями не имеет практических последствий: два выпуска облигаций имеют разные купонные ставки, так как они выпущены в разное время, но со временем их рыночная цена изменилась таким образом, что доходность стала одинаковой.

Средневзвешенное стоимости акционерного капитала можно объяснить следующим образом. Если Leveraged Industries, Inc. решит увеличить добавочный капитал без изменения своей основной бизнес-стратегии (т.е. типа проектов, риска проектов) и сохранить тот же набор ценных бумаг, то необходимо будет выплачивать в год 12,9%, или $129 на $1000. Эти $129 будут выплачены по различным типам ценных бумаг в соответствии с их весами.

Пример. *Корректировка недостаточной репрезентативности*

Кроме того, взвешенное среднее используют, чтобы скорректировать недостатки репрезентативности выборки по отношению к интересующей вас генеральной совокупности. Поскольку среднее выборки учитывает все элементы одинаково, а вам известно, что (по сравнению с генеральной совокупностью) некоторые группы элементов представлены избыточно, а другие, наоборот, — недостаточно, то более точный результат можно получить, используя взвешенное среднее. Взвешенное среднее будет точнее, поскольку

[6] Например, первый вес равен 100 000/455 000 = 0,220. Сумма весов не равна 1 из-за ошибки при округлении. Это не беда. Если вам нужна большая точность, можно увеличить количество значащих цифр после запятой до четырех, пяти и более, пока вы не получите результат с необходимой точностью.

в нем известная информация о каждой группе (взятая из выборки) будет объединена с дополнительной информацией о представительстве каждой группы (в генеральной совокупности, а не в выборке).

Снова рассмотрим выборку 300 жителей Кливленда, которую мы анализировали ранее с точки зрения затрат людей на медицинские товары. Предположим, что процент молодых людей (до 18 лет) в этой выборке (21,7%) не соответствует известному проценту для всего населения города (25,8%) и что средние денежные расходы, подсчитанные для каждой группы отдельно, составляют:

средние денежные расходы для людей моложе 18 лет — $4,86;

средние денежные расходы для людей старше 18 лет — $7,06.

При вычислении средневзвешенного этих затрат будем использовать веса не выборки, а известные нам веса генеральной совокупности, т.е. будем считать, что имеем дело с 25,8% молодых людей и 74,2% людей старше 18 лет (разность 100% — 25,8%). Конечно, если бы были известны оценки расходов для города в целом, то вы бы их также использовали. Но такие данные вам недоступны. Вам известны расходы только для 300 человек из выборки. После преобразования процентов в веса взвешенное среднее вычисляется следующим образом:

взвешенное среднее расходов = (0,258 x $4,86)+(0,742 x $7,08) = $6,49.

Взвешенное среднее $6,49 дает лучшую оценку среднего значения расходов на медицинские товары в Кливленде, чем обычное (не взвешенное) среднее ($6,58). Взвешенное среднее лучше, поскольку оно содержит *поправку* на слишком большой процент людей в возрасте старше 18 лет в нашей выборке из 300 человек[7]. Так как люди такого возраста тратят больше, то без поправки средняя оценка расходов получается завышенной ($6,58 по сравнению с $6,49).

Конечно же, даже эта новая взвешенная оценка может быть неверной. Но она основана на большем объеме информации, поэтому ожидаемая ошибка будет меньше, что можно доказать с помощью математических моделей. Новая оценка не обязательно каждый раз будет лучше (т.е. и в данном примере обычное, не взвешенное, среднее может в действительности быть ближе к истине), но *вероятность* того, что взвешенная оценка будет ближе к истине, *намного больше*.

Медиана: типическое значение для количественных и порядковых данных

Медиана — это значение, которое расположено посередине; половина элементов в наборе данных больше этого значения, а вторая половина — меньше. Таким образом, медиана располагается в *центре* данных и дает представление о списке значений. Чтобы найти медиану, данные располагают в порядке возрастания, а затем определяют среднее значение. Обратите внимание, что если в наборе данных нет одного центрального значения, то следует усреднить те два значения, которые расположены посередине ряда.

Медиану можно определить в терминах *рангов*[8]. **Ранги** связывают числа 1, 2, 3, ... n со значениями данных таким образом, что наименьшее значение имеет ранг 1, следующее по величине значение — ранг 2 и так далее до наибольшего значения, которое имеет ранг n. В основу определения медианы положен следующий принцип.

[7] Статистики часто говорят о "поправке" или "внесении корректив" с учетом того или иного фактора. Здесь представлен один из способов; позже вы узнаете о множественной регрессии — другом мощном средстве учета влияния факторов, которые не находятся под вашим контролем.

[8] Ранги образуют основу *непараметрических методов*, которые будут описаны в главе 16.

Ранг медианы

Медиана имеет ранг $(1 + n)/2$.

С учетом всех возможных особых случаев медиана для списка из n элементов вычисляется таким образом.

1. Расположите элементы данных в порядке возрастания (или уменьшения — это не имеет значения).
2. Определите среднее значение полученного ряда. Возможны два варианта.

   а) Если n — нечетное число, то медианой будет среднее значение данных, которое имеет номер $(1+n)/2$, если отсчитывать от любого из двух концов упорядоченного списка. Например, медиана списка 15, 27, 14, 18, 21 из $n = 5$ значений равняется:

   медиана (15, 27, 14, 18, 21) = медиана (14, 15, 18, 21, 27) = 18.

   Следует отметить, что медиана, 18, это третье по порядку значение в упорядоченном списке, что соответствует формуле, поскольку $(1+n)/2 = (1 + 5)/2 = 3$.

   В качестве примера порядковых данных рассмотрим список рейтингов облигаций AAA, A, B, AA, A. Для этого списка медиана будет вычисляться следующим образом:

   медиана (AAA, A, B, AA, A) = медиана (B, A, A, AA, AAA) = A.

   б) Если n — четное число, то ряд имеет не одно, а два средних значения. Эти значения расположены на расстоянии $(1+n)/2$ от каждого из двух концов упорядоченного списка данных.

   в) Если набор данных *количественный* (т.е. состоит из чисел), то медианой является среднее этих двух значений, расположенных в середине ряда. Например, медиана списка 15, 27, 14, 18 из $n = 4$ чисел вычисляется следующим образом:

   медиана (15, 27, 14, 18) = медиана (14, 15, 18, 27) = (15 + 18)/2 = 16,5.

   В этом случае по формуле $(1+n)/2$ имеем: (1 + 4)/2 = 2,5; что говорит о необходимости пройти в упорядоченном списке половину пути между вторым и третьим числом, усреднив эти два числа.

   г) Если набор данных является *порядковым* (т.е. содержит упорядоченные категории) и если два расположенных в середине ряда значения представляют одну и ту же категорию, то эта категория является медианой. Если эти два значения представляют различные категории, то обе эти категории будут медианами. Например, для списка рейтингов облигаций A, B, AA, A медиана будет равна:

   медиана (A, B, AA, A) = медиана (B, A, A, AA) = A,

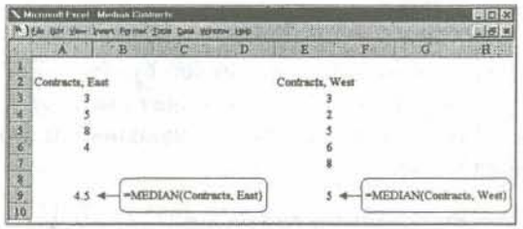
   поскольку оба расположенных посередине значения равны A.

   В другом примере для списка рейтингов облигаций A, AAA, B, AA, AAA, B медиана будет вычисляться следующим образом:

медиана (A, AAA, B, AA, AAA) = медиана (B, B, A, AA, AAA, AAA) =
= A и AA.

Это лучшее, что можно сделать в данной ситуации, так как для порядковых данных нельзя вычислить среднее двух значений.

Для вычисления медианы в Excel можно использовать функцию MEDIAN (МЕДИАНА) следующим образом.



Чем отличается медиана от среднего? Если набор данных распределен нормально, то значения медианы и среднего близки между собой, поскольку нормальное распределение симметрично и имеет четко выраженную среднюю точку. Однако даже при нормальном распределении (здесь речь идет о "практически нормальном" распределении, а не о теоретически нормальном распределении. — *Прим. ред.*) среднее и медиана несколько отличаются друг от друга, поскольку каждая из этих величин определяется по-своему и, кроме того, в реальных данных почти всегда присутствует некоторая случайность. Если набор данных не подчиняется нормальному распределению, то медиана и среднее могут сильно различаться, потому что у асимметричного распределения нет четко выраженной центральной точки. Обычно среднее по отношению к медиане сдвинуто в направлении более длинного хвоста или в направлении выброса, поскольку среднее реально учитывает значения таких экстремальных наблюдений, в то время как для медианы важно лишь, по какую сторону от нее лежит то или иное значение.

Пример. *Обвал фондового рынка 19 октября 1987 года: падение акций в первый день*

Обвал фондового рынка 1987 года стал экстраординарным событием, когда рынок потерял за один день 20% стоимости. В этом примере мы определим объем потерь в первый день кризиса биржи.

Рассмотрим процент потерь стоимости акций 29 компаний из списка Dow Industrial в промежуток времени между закрытием торгов в пятницу 16 октября и открытием торгов в понедельник 19 октября 1987 года, в день краха. Из табл. 4.1.1 видно, что даже при открытии торгов эти акции уже потеряли значительную часть своей стоимости.

Гистограмма на рис. 4.1.2 показывает, что распределение достаточно близко к нормальному.

Наблюдается небольшая асимметрия в направлении низких значений (т.е. хвост слева слегка длиннее, чем справа), но несмотря на это, распределение приблизительно нормальное со случайными отклонениями. Среднее значение процентного изменения –8,2% и медиана процентного изменения –8,6% довольно близки друг к другу. Действительно, гистограмма имеет четко выраженную центральную область, поэтому любая разумная обобщающая характеристика должна быть расположена вблизи этой центральной области.

Таблица 4.1.1. Падение акций при открытии торгов в день обвала фондового рынка 1987 года

Фирма	Изменение стоимости, %	Фирма	Изменение стоимости, %
Union Carbide	–4,1	Primerica	–6,8
USX	–5,1	Navistar	–2,1
Bethlehem Steel	–4,5	General Electric	–17,2
AT&T	–5,4	Westinghouse	–15,7
Boeing	–4,0	Alcoa	–8,9
International Paper	–11,6	Kodak	–15,7
Chevron	–4,0	Texaco	–12,3
Woolworth	–3,0	IBM	–9,6
United Technologies	–4,4	Merck	–12,0
Allied-Signal	–9,3	Phillip Morris	–12,4
General Motors	–0,9	Du Pont	–8,6
Procter & Gamble	–3,5	Sears Roebuck	–11,4
Coca-Cola	–10,5	Goodyear Tire	–10,9
McDonald's	–7,2	Exxon	–8,6
Minnesota Mining	–8,9		

Данные взяты из статьи "Trading in the 30 Dow Industrials Shows Wide Damage of Oct. 19 Crash", *The Wall Street Journal*, 1987, December, 16 p. 20. Источник включает данные только о 29 указанных ценных бумагах. Отрицательные числа свидетельствуют о падении стоимости. Указанные для всех 29 компаний отрицательные числа свидетельствует о том, что на момент открытия биржи акции всех компаний упали в цене по сравнению с предыдущим днем торгов.

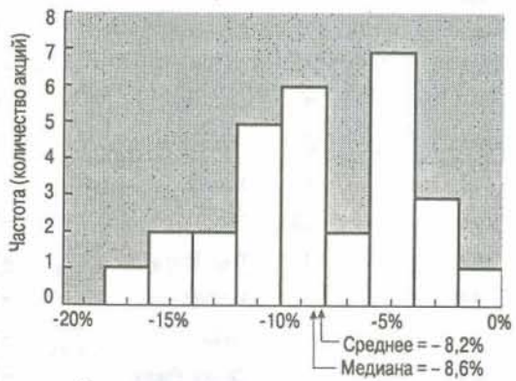


Рис. 4.1.2. Распределение значений процента падения стоимости акций 29 промышленных компаний на момент открытия торгов в день обвала 19 октября 1987 года

Среднее процентное изменение –8,2% можно интерпретировать следующим образом. Если в пятницу на момент закрытия торгов у вас был портфель инвестиций с одинаковым количеством денег, вложенных в каждую из этих ценных бумаг (в соответствии со стоимостью акций на момент закрытия торгов в пятницу), то в понедельник при продаже в начале торгов ваш инвестиционный портфель потерял бы 8,2% своей стоимости. Разные акции потеряли разные проценты своей стоимости, но портфель инвестиций с одинаковыми весами акций потерял бы именно это среднее значение стоимости. А что если вы вложили разное количество средств в разные акции? Тогда потерю стоимости инвестиционного портфеля можно было бы рассчитать как средневзвешенное, используя для определения весов размеры вложенных средств.

Медиану процентного изменения –8,6% можно интерпретировать следующим образом. Если упорядочить значения процентных изменений стоимости акций, то половина акций потеряла в стоимости 8,6% или больше, а половина акций потеряла в стоимости 8,6% или меньше. Таким образом, падение стоимости акций на 8,6% представляет собой некий средний результат для этой группы акций. В табл. 4.1.2 содержится упорядоченный список. Обратите внимание, что Exxon находится в центре списка на 15-м месте, поскольку (29 + 1) = 15.

Всякое падение акций обращает на себя внимание. Но падение более чем на 8% в начале торгов является угрожающим сигналом. В тот день среднее падение индекса Dow Jones Industrial было рекордным — 508 пунктов, или 22,6%. Это стало подлинной трагедией для многих людей и организаций[9].

Таблица 4.1.2. Список акций, упорядоченный по размеру потерь на момент открытия торгов при обвале фондового рынка в 1987 году

Фирма	Изменение стоимости, %	Ранг	Фирма	Изменение стоимости, %	Ранг
General Motors	–0,9	1	Minnesota Mining	–8,9	16
Navistar	–2,1	2	Alcoa	–8,9	17
Woolworth	–3,0	3	Allied-Signal	–9,3	18
Procter & Gamble	–3,5	4	IBM	–6,9	19
Boeing	–4,0	5	Coca-Cola	–10,5	20
Chevron	–4,0	6	Goodyear Tire	–10,9	21
Union Carbide	–4,1	7	Sears Roebuck	–11,4	22
United Technologies	–4,4	8	International Paper	–11,6	23
Bethlehem Steel	–4,5	9	Merck	–12,0	24
USX	–5,1	10	Texaco	–12,3	25
AT&T	–5,4	11	Philip Morris	–12,4	26
Primerica	–6,8	12	Westinghouse	–15,7	27
McDonald's	–7,2	13	Kodak	–15,7	28
Du Pont	–8,6	14	General Electric	–17,2	29
Exxon	–8,6	15			

[9] Значение 22,6% взято из *The Wall Street Journal*, 1987, October, 20, p. 1.

Пример. *Личные доходы*

Распределение таких количественных данных, как личные доходы отдельных людей и семей (как и распределение продаж, трат, цен и т.п.), часто скошено в сторону более высоких значений, поскольку такие наборы данных содержат много небольших значений, некоторое количество средних значений и немного больших и очень больших значений. Таким образом, обычно среднее больше, чем медиана. Это связано с тем, что на значение среднего, получаемого сложением всех элементов, сильно влияют большие значения. Рассмотрим доходы домохозяйств в США в 1992 году[10]:

среднее доходов домохозяйств — $ 39 020;

медиана доходов домохозяйств — $ 30 786.

Среднее дохода выше, чем медиана, потому что на значение среднего оказывают сильное влияние относительно небольшое количество очень высокодоходных домохозяйств. Вспомним, что при вычислении среднего эти высокие доходы входят в сумму, а при вычислении медианы они являются просто "высокими доходами" (при этом каждому домохозяйству с высокими доходами соответствует домохозяйство с низкими доходами).

Гистограмма на рис. 4.1.3 показывает вид распределения доходов для выборки из 100 человек.

Распределение сильно скошено в направлении высоких доходов, поскольку есть много людей с низкими доходами (на это указывают высокие столбики слева на гистограмме) и относительно немного людей, имеющих средние и высокие доходы (короткие столбики в середине и справа на гистограмме). Среднее значение дохода $38 710 выше, чем медиана $27 216. Медиана (точка, которая делит количество объектов пополам) ниже среднего, потому что на данной гистограмме большинство людей имеют низкие доходы, а наличие людей с высокими доходами значительно увеличивает значение среднего.

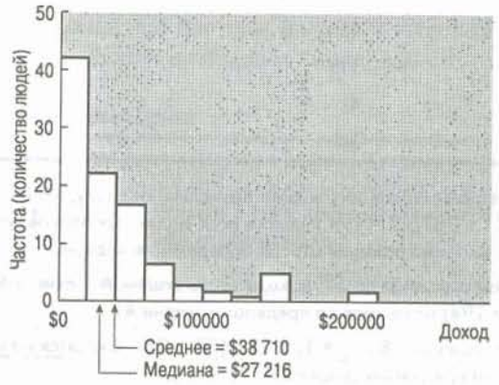


Рис. 4.1.3. Гистограмма распределения данных о доходах 100 человек. Это асимметричное распределение и среднее значительно больше, чем медиана

[10] Данные *Statistical Abstract of the United States*: 1994 (114 th ed.), Washington, D.C., 1994, p. 465, 466.

Пример. *Стадии сборки компьютерных системных блоков*

Рассмотрим процесс производства компьютеров, состоящий из следующих стадий.

Производство материнской платы.

Установка разъемов на материнскую плату.

Установка в разъемы электронных микросхем.

Тестирование собранной материнской платы.

Установка собранной материнской платы в системный блок компьютера.

Тестирование собранного системного блока.

Если у вас имеется набор данных, в котором для каждого системного блока указано, на какой из производственных стадий изготовления он находится, то такой одномерный набор порядковых данных может иметь следующий вид:

A, C, E, F, C, C, D, C, A, E, E, ...

Этот набор данных является порядковым, поскольку для категорий существует естественный порядок — порядок прохождения изделия через все стадии производственного процесса от начала сборки до завершения. Такой набор данных можно представить в виде списка частот следующего вида.

Стадия производства	Количество компьютерных системных блоков
A	57
B	38
C	86
D	45
E	119
F	42
Итого	387

Поскольку это порядковые данные, для них можно вычислить медиану, но не среднее. Медианой будет системный блок с рангом (1 + 387)/2 = 194 в списке всех системных блоков, упорядоченных в соответствии со стадией производства. Ниже показан способ определения медианы.

Блоки с рангами от 1 до 57 находятся на стадии A. Таким образом, медиана (которая имеет ранг 194) находится за пределами стадии A.

Блоки с рангами от 58 (57 + 1) до 95 (57 + 38) находятся на стадии B. Значит, медиана находится за пределами стадии B.

Блоки с рангами от 96 (95 + 1) до 181 (95 + 86) находятся на стадии C. Следовательно, медиана находится за пределами стадии C.

Блоки с рангами от 182 (181 + 1) до 226 (181 + 45) находятся на стадии D. Таким образом, медиана находится на стадии D, поскольку ранг медианы (194) лежит между рангами 182 и 226.

Таким образом, около половины системных блоков находятся на стадиях, предшествующих стадии D, и примерно половина — на стадиях, следующих за стадией D. Поэтому стадия D является средней точкой (с точки зрения готовности сборки) для всех системных блоков, находящихся в настоящий момент в производстве.

Мода: типическое значение даже для номинальных данных

Мода представляет собой наиболее распространенную категорию, т.е. категорию, которая чаще всего встречается в наборе данных. Это единственная характеристика, которую можно определить для номинальных качественных данных, поскольку неупорядоченные категории нельзя складывать (как это требуется для среднего) и нельзя ранжировать (как это требуется для медианы). Моду можно легко найти для порядковых данных, если просто проигнорировать упорядоченность категорий и выполнять все действия так же, как для набора номинальных данных с неупорядоченными категориями.

Мода также определена для количественных данных (чисел), хотя при этом может иметь место некоторая неопределенность. Для количественных данных моду можно определить как значение, соответствующее наивысшей точке на гистограмме, возможно, на середине самого высокого столбика. Источники неопределенности могут быть разными. На гистограмме может быть два "самых высоких" столбика. Или, что значительно хуже, определение моды может зависеть от того, каким образом построена диаграмма: изменение ширины столбиков и их расположения может привести к небольшим (или умеренным) изменениям формы распределения, в результате чего может измениться и мода. Для количественных данных мода является несколько неопределенным понятием.

Моду найти легко. Независимо от того, представляют имеющиеся у вас числа количество объектов в каждой категории или соответствующие проценты, необходимо просто выбрать категорию с самым большим количеством или процентом. Если на первое место претендуют две или больше категорий, то необходимо указать все эти категории под общим названием "мода" для этого набора данных.

Пример. *Голосование на выборах*

Поскольку во время выборов подсчитывают количество отданных голосов, то эти голоса можно рассматривать как набор номинальных качественных данных. У вас может быть свое мнение относительно упорядочения кандидатов, но так как общего согласия в этом вопросе нет, то вы можете считать этот набор данных неупорядоченным. Список данных может выглядеть так:

Смит, Джонс, Баттерсвоурт, Смит, Смит, Баттерсвоурт, Смит...

Результаты выборов можно записать следующим образом.

Фамилия	Количество голосов	Процент
Баттерсвоурт	7175	15,1
Джонс	18 956	39,9
Харвей	502	1,1
Смит	20 817	43,9
Итого	47450	100,0

Ясно, что модой в этом наборе данных будет Смит, поскольку он набрал наибольшее количество голосов (20 817) и наибольший процент голосов (43,9%). Обратите внимание, что мода не обязательно представляет больше половины (большинство) объектов, хотя иногда может быть и так. Мода просто представляет больше объектов, чем любая другая категория.

Пример. *Контроль качества: отклонения в производстве*

Важным видом деятельности при создании качественных изделий является анализ отклонений в производственных процессах. Одни отклонения от производственного процесса неизбежны, но допустимы (из-за небольшой величины), в то время как другие выводят процесс из-под контроля и приводят к производству низкосортных изделий. Контроль качества будет подробно рассмотрен в главе 18. Эдвардс Деминг (W. Edwards Deming) впервые ввел контроль качества в Японии в 50-е годы. Некоторые из его методов кратко можно обобщить следующим образом.

Предложенный Демингом метод в основе своей является статистическим. Любая производственная деятельность, в цеху или в офисе, имеет отклонения от идеала. Деминг предложил систематический метод измерения отклонений производственного процесса, выявления причин этих отклонений и их уменьшения, совершенствования за счет этого процесса, а значит, и повышения качества продукции[11].

Сбор и последующий анализ данных — это ключевой компонент хорошего контроля качества. Предположим, что предприятие регистрирует причину брака каждый раз при появлении изделия недопустимого качества.

Причина проблемы	Число случаев
Пайка соединений	37
Пластмассовый корпус	86
Блок питания	194
Грязь	8
Удар (при падении)	1

Ясно, что модой в этом наборе данных является блок питания, поскольку эта причина брака встречается чаще других. Мода помогает сосредоточить внимание на самой важной категории (наиболее часто встречающейся). Нет необходимости разрабатывать дополнительные мероприятия по поддержанию чистоты на рабочем месте или по недопущению падения коробок, поскольку эти причины мало влияют на общую частоту брака. В первую очередь следует обратить внимание на модальную категорию.

В рассмотренной ситуации фирма могла бы попробовать разобраться с проблемой "блока питания" и принять соответствующие меры. Возможно, этот блок питания имеет недостаточную мощность для данного изделия и необходим более мощный источник. Возможно, нужно найти более надежного поставщика. В любом случае мода помогает уточнить имеющуюся проблему.

Пример. *Повторное рассмотрение стадий сборки системных блоков*

Рассмотрим еще раз описанный раньше пример данных о состоянии сборки системных компьютерных блоков. Ниже приведен набор данных.

Стадия производства	Количество системных блоков
A	57
B	38
C	86
D	45
E	119
F	42
Итого	387

[11] "The Curmudgeon Who Talks Tough on Quality", *Fortune*, 1984, June, 25, p. 119.

Раньше мы уже определили, что медиана приходится на стадию производства D, поскольку эта стадия отделяет половину системных блоков, находящихся на начальных стадиях сборки, от второй половины системных блоков на конечной стадии сборки. Однако в данном случае медиана не совпадает с модой (хотя в некоторых других примерах мода может совпадать с медианой).

Здесь мода представляет собой стадию E, на которой находится 119 системных блоков, т.е. больше, чем на любой другой стадии. В такой ситуации руководство должно быть проинформировано о моде, потому что наиболее "узкое место" в производственном процессе, скорее всего, проявится именно как мода.

В рассмотренном примере стадия E — это установка материнской платы в системный блок. Наличие большого количества системных блоков на этой стадии может быть связано с большей трудоемкостью данной операции. Но, с другой стороны, это может быть и свидетельством наличия проблем у служащих, работающих на этой стадии (возможно, причина в недостаточном количестве людей или большом количестве отсутствующих работников). В таком случае руководству необходимо обратить на это внимание.

Какие показатели нужно использовать

Какой их трех показателей (среднее, медиану или моду) следует использовать в конкретных обстоятельствах? Есть два вида ответов. Первый зависит от того, что можно вычислить, а второй зависит от того, какой из показателей более полезен.

Моду можно вычислить для любого одномерного набора данных (хотя в случае количественных данных проблемой может быть некоторая неопределенность). Среднее можно вычислить только для количественных данных (чисел), а медиану — для всех типов данных, кроме номинальных (неупорядоченных категорий). Таким образом, ваш выбор ограничен, а в случае номинальных данных у вас вообще нет другого выбора, кроме как использовать моду. Рекомендации по выбору характеристики в зависимости от типа данных можно представить таким образом.

	Количественные	Порядковые	Номинальные
Среднее	Да		
Медиана	Да	Да	
Мода	Да	Да	Да

В случае количественных данных, для которых можно вычислить все три характеристики, насколько они отличаются между собой? Если распределение близко к нормальному, разница невелика, поскольку каждая из характеристик стремится к четко выраженной середине, имеющей форму колокола кривой распределения (рис. 4.1.4).

Однако в случае асимметричного распределения данных эти характеристики могут заметно различаться (как мы уже отмечали для среднего и медианы). На рис. 4.1.5 показаны рассматриваемые характеристики для данных, не подчиняющихся нормальному распределению.

Среднее следует использовать, когда набор данных распределен нормально (по крайней мере приблизительно), поскольку в этом случае среднее является самой эффективной характеристикой. Среднее также следует вычислять и в тех ситуациях, где необходимо сохранить или предсказать общую сумму значений данных, так как другие характеристики не позволяют это сделать.

Медиана служит хорошей характеристикой асимметричного распределения, поскольку на него не влияет небольшое число данных с высокими значениями.

В случае сильной асимметрии медиана значительно лучше среднего характеризует большинство данных. Медиана также полезна при наличии выбросов значений, так как она устойчива к их влиянию. Медиана полезна для порядковых данных (упорядоченные категории), хотя в зависимости от решаемого вопроса можно использовать и моду.

Моду используют при наличии номинальных данных, так как в этом случае нельзя вычислять среднее и медиану. Она также полезна для порядковых данных, когда важно определить наиболее распространенную категорию.

Помимо рассмотренных существует много других характеристик. Перспективным является использование так называемых "робастных" (устойчивых) оценок,

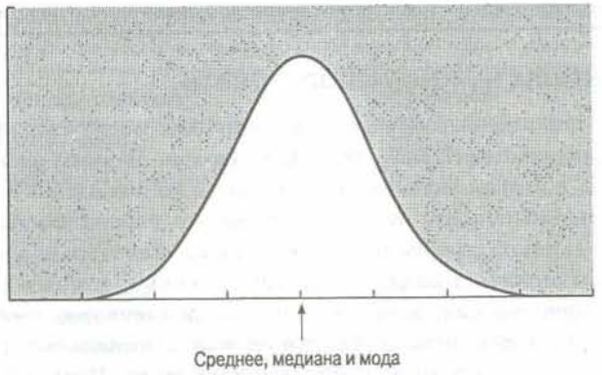


Рис. 4.1.4. Для идеального нормального распределения среднее, медиана и мода совпадают. Для реальных данных, где всегда присутствует случайность, эти характеристики будут приблизительно, но не точно, равны между собой

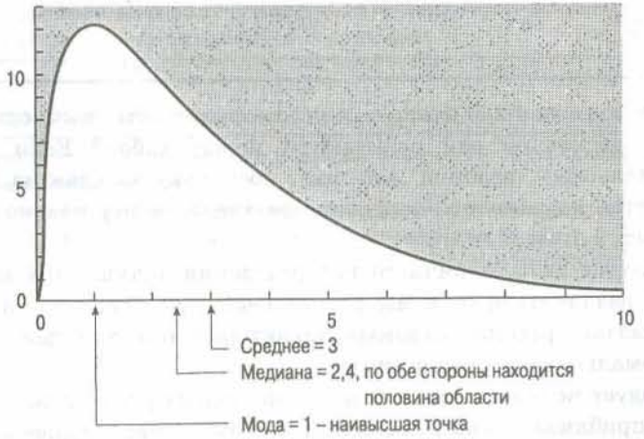


Рис. 4.1.5. Для скошенного распределения среднее, медиана и мода различаются. Мода соответствует наивысшей точке на кривой распределения. По обе стороны от медианы находится половина области под кривой распределения. Среднее находится в точке центра тяжести распределения, как точки опоры доски детских качелей

которые сочетают в себе лучшие свойства среднего и медианы[12]. Для нормально распределенных данных такие оценки представляют достаточно эффективный выбор и в то же время они, как и медиана, устойчивы к влиянию выбросов.

4.2. Что такое перцентиль

Перцентили — это характеристики набора данных, которые выражают ранги элементов в виде процентов от 0 до 100%, а не в виде чисел от 1 до n, таким образом, что наименьшему значению соответствует нулевой перцентиль, наибольшему — 100-й перцентиль, медиане — 50-й перцентиль и т.д. Перцентили можно рассматривать как показатели, разбивающие наборы количественных и порядковых данных на определенные части.

Обратите внимание, что перцентиль представляет собой имеющий определенный ранг элемент данных и выражен в тех же единицах, что и единицы набора данных. Например, 60-й перцентиль эффективности продаж может быть равен $385 062 (измерен *не* в процентах, а в долларах, как и элементы набора данных). Если этот 60-й перцентиль, равный $385 062, характеризует деятельность определенного агента по продажам (например, Мари), то это означает, что приблизительно 60% других агентов имеют результаты ниже, чем у Мари, а 40% агентов имеют более высокие результаты.

Перцентили используют для двух целей.

1. Чтобы показать значение элемента в данных при заданном перцентильном ранге (например, "10-й перцентиль равен $156 293 ").
2. Чтобы показать перцентильный ранг значения данного элемента в наборе данных (например, "эффективность продаж агента по сбыту (Джона) составляет $296 994, что соответствует 55-му перцентилю").

Экстремумы, квартили и блочные диаграммы

Перцентили играют важную роль в качестве опорных характеристик. Чтобы обобщить основные черты распределения, достаточно нескольких значений перцентилей. Так, 50-й перцентиль — это медиана, поскольку 50-й перцентиль находится посередине между наибольшим и наименьшим значениями ряда. Интерес представляют **экстремумы** — *наибольшее* и *наименьшее* значения данных, т.е. 0-й и 100-й перцентили соответственно. Дополняют набор базовых характеристик **квартили**, определяемые как 25-й и 75-й перцентили.

Удивительно, но статистики до сих пор спорят относительно точного определения квартилей, поскольку их можно вычислять разными способами. Идея квартилей понятна. Квартили — это значения ранжированного ряда, которые находятся на расстоянии одной четвертой на пути от наименьшего и наибольшего значений. Однако эта формулировка не указывает точно, как вычислять квар-

[12] Более подробную информацию об устойчивых оценках можно найти в книге Hoaglin D.C., Mosteller F., and Tukey J. W. *Understanding Robust and Explanatory Data Analysis.* — New York: Wiley, 1983.

тили. Джон Тьюки, один из создателей практического анализа данных, определяет квартили таким образом[13].

1. Вычисляем ранг медианы по формуле $(1+n)/2$ и отбрасываем дробную часть. Например, при $n=13$ получаем $(1+13)/2=7$. При $n=24$ отбрасываем дробную часть у $(1+24)/2=12{,}5$ и получаем 12.
2. Добавляем к полученному значению 1 и делим на 2. Полученное значение представляет собой *ранг нижнего квартиля*. Например, при $n=13$ ранг нижнего квартиля равен $(1+7)/2=4$. При $n=24$ ранг нижнего квартиля равен $(1+12)/2=6{,}5$, что свидетельствует о необходимости усреднить значения с рангами 6 и 7.
3. Отнимаем полученное значение от $(n+1)$. Результатом будет *ранг верхнего квартиля*. Например, при $n=13$ получим $(13+1)-4=10$. При $n=24$ получаем $(1+24)-6{,}5=18{,}5$, что свидетельствует о необходимости усреднить значения с рангами 18 и 19.

Значения квартилей находят исходя из этих рангов. Ниже приведена общая формула определения рангов квартилей, которая представляет указанные выше шаги вычислений.

Ранги квартилей

$$\text{Ранг нижнего квартиля} = \frac{1+\text{int}[(1+n)/2]}{2};$$

$$\text{Ранг верхнего квартиля} = n+1-\text{Ранг нижнего квартиля},$$

где int означает функцию взятия целого, которая отбрасывает дробную часть числа.

Пять базовых показателей включают наименьшее значение, нижний квартиль, медиану, верхний квартиль, наибольшее значение.

Пять базовых показателей

Наименьшее значение данных (0-й перцентиль).

Нижний квартиль (25-й перцентиль, на четверть расстояния от наименьшего значения).

Медиана (50-й перцентиль, середина).

Верхний квартиль (75-й перцентиль, на три четверти расстояния от наименьшего значения или на четверть расстояния от наибольшего значения).

Наибольшее значение (100-й перцентиль).

Вместе эти характеристики дают достаточно ясное представление об особенностях еще не обработанного набора данных. Два экстремума характеризуют размах (диапазон) данных, медиана показывает центр, два квартиля определяют границы, "расположенной в центре половины данных", а положение медианы относительно квартилей дает грубое представление о наличии или отсутствии асимметрии.

[13] Tukey J. W. *Exploratory Data Analysis.* — Reading, Mass.: Addison-Wesley, 1977. Тьюки рассматривает квартили как опоры и дает соответствующее определение на с. 33. Функция вычисления квартилей в Excel может дать несколько иные значения, поскольку иногда для вычислений используется взвешенное среднее.

Блочная диаграмма — это изображение всех пяти указанных показателей (рис. 4.2.1).

Блочная диаграмма, как и гистограмма, дает визуальное представление о распределении, но использует иной способ графического отображения. Блочная диаграмма не содержит мелких деталей, что позволяет охватить всю картину в целом и сравнивать несколько групп чисел, не вдаваясь в детали каждой из групп. При необходимости подробно рассмотреть форму распределения лучше использовать гистограмму.

Подробная блочная диаграмма — это блочная диаграмма, которая также содержит помеченные метками выбросы (метки также используют для показа экстремальных наблюдений, не являющихся выбросами). Метки выделяют те наблюдения, которые требуют особого внимания. При создании подробной блочной диаграммы **выбросы** определяют как те значения данных (если они есть), которые расположены далеко от центра распределения. В частности, большое значение в наборе данных рассматривается как выброс, если оно превышает

верхний квартиль + 1,5 (верхний квартиль – нижний квартиль).

Малое значение в наборе данных рассматривается как выброс, если оно меньше, чем

нижний квартиль – 1,5 (верхний квартиль – нижний квартиль).

Так выбросы определяет Тьюки[14]. В дополнение к нанесению на диаграмму выбросов с соответствующими метками можно также отметить экстремальные значения, которые выбросами *не* являются (по одному с каждой стороны), поскольку часто они также заслуживают особого внимания. На рис. 4.2.2 для сравнения показаны блочная и подробная блочная диаграммы.

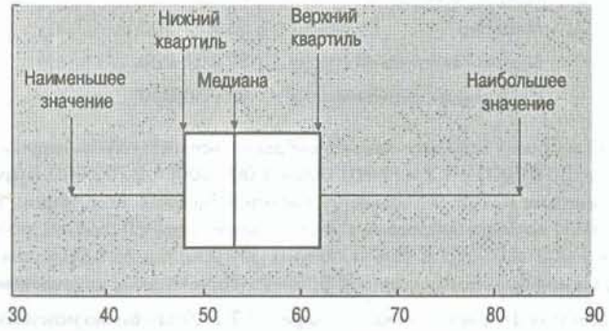


Рис. 4.2.1. Блочная диаграмма содержит пять базовых показателей одномерного набора данных и позволяет быстро определить характер распределения

[14] Tukey J.W. *Exploratory Data Analysis*, p. 44. Также см. главу 3 книги Hoaglin et al. *Understanding Robust and Exploratory Data Analysis.*

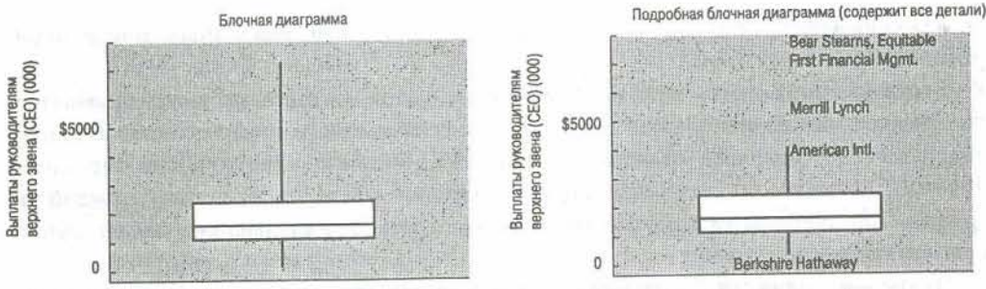


Рис. 4.2.2. Блочная диаграмма (слева) и подробная блочная диаграмма (справа) вознаграждения руководителям финансовых компаний. Обе диаграммы содержат пять базовых показателей, но подробная блочная диаграмма дает ценную информацию о выбросах (а также показывает экстремальные значения, которые выбросами не являются). В этом примере выбросы представляют собой фирмы, выплатившие исключительно высокое вознаграждение своим руководителям

Пример. Выплаты руководителям

Рассмотрим выплаты (заработная плата и премии) руководителям финансовых компаний в 1994 году. Табл. 4.2.1 содержит упорядоченный список размеров выплат, их ранги и соответствующие пять характеристик распределения.

Таблица содержит данные о $n = 38$ фирмах, следовательно, медиана ($1 497 500) имеет ранг (1+38)/2= =19,5 и представляет собой среднее значение выплат руководителям фирм Transamerica (ранг 19) и General RE (ранг 20). Нижний квартиль ($1 000 000) имеет ранг (1 + 19)/2 = 10 и представляет собой выплаты, полученные R.E. Denham из фирмы Salomon. Верхний квартиль ($2 101 000) имеет ранг 38+1–10 = 29 и представляет собой выплаты A.J.C. Smith из фирмы March & McLennan. Ниже приведены пять базовых показателей для набора данных о размерах выплат руководителям этих 38 финансовых фирм.

Наименьшее значение	$100 000
Нижний квартиль	$1 000 000
Медиана	$1 497 500
Верхний квартиль	$2 101 000
Наибольшее значение	$7 730 000

Есть ли среди значений выбросы? Если рассчитывать выбросы с использованием квартилей, то выплаты, размер которых превышает 2 101 000 + 1,5 × (2 101 000 – 1 000 000) = 3 752 000, будут выбросами. Таким образом, пять самых высоких выплат (выплаченные фирмами Equitable, Bear Stearns, First Financial Mgmt., Merrill Lynch, and Travelers) являются выбросами в верхней части. С другой стороны, любые выплаты, размер которых меньше, чем 1 000 000 – 1,5 × (2 101 000 – 1 000 000) = – 651 500, также будут выбросами. Поскольку размер наименьшей выплаты равен 100 000, то в нижней части распределения выбросов нет.

Блочные диаграммы для этих 38 фирм приведены на рис. 4.2.2. Хотя обычно используют одну диаграмму (вероятно, с большим количеством подробностей), мы для сравнения приводим здесь обе диаграммы.

Одно из преимуществ блочных диаграмм заключается в том, что они позволяют сконцентрировать внимание на основных особенностях нескольких наборов данных одновременно, не отвлекаясь на детали. Рассмотрим выплаты, полученные в 1994 году руководителями крупных банков, предприятий фармацевтической отрасли, коммунальных предприятий и финансовых компаний[15]. Теперь мы имеем четыре самостоятельных набора данных: по одному одномерному набору данных (набору значений) для каждой из

[15] Данные взяты из "Executive Compensation Scoreboard". *Business Week*, 1995, April, 24, p. 94–119.

четырех отраслей. Это означает, что для каждой из отраслей можно вычислить пять основных показателей и построить блочную диаграмму.

Расположив построенные в одном масштабе блочные диаграммы на одном рисунке (рис. 4.2.3), можно легко сравнить между собой типичные размеры выплат руководителям в разных отраслях[16].

Таблица 4.2.1. Выплаты руководителям финансовых компаний

Фирма	Руководитель	Зарплата и премии в 1994 году, дол.	Ранг
Equitable*	R. H. Jenrette	7 730 000	38 Наибольшее значение равно $7 730 000
Bear Stearns*	J. E. Cayne	7 666 000	37
First Financial Mgmt.*	P. H. Thomas	6 910 000	36
Merrill Lynch*	D. P. Tully	4 840 000	35
Travelers*	S. I. Weill	3 903 000	34
American Intl. Group	M. R. Greenberg	3 750 000	33
Schwab (Charles)	C. R. Schwab	3 273 000	32
Dean Witter Discover	P. J. Purcell	3 200 000	31
American Express	H. Golub	3 077 000	30
Marsh & McLennan	A. J. Smith	2 101 000	29 Верхний квартиль = $2 101 000
Progressive	P. B. Lewis	2 063 000	28
American General	H. S. Hook	1 960 000	27
Loews	P. R. Tisch	1 937 000	26
Torchmark	R. K. Richey	1 936 000	25
Household International	D. C. Clark	1 877 000	24
Aflac	D. P. Amos	1 726 000	23
Cigna	W. H. Taylor	1 723 000	22
Great Western Financial	J. F. Montgomery	1 674 000	21
Transamerica	F. C. Herringer	1 537 000	20
			Медиана = $1 497 500
General RE	R. E. Ferguson	1 458 000	19
Chubb	D. R. O'Hare	1 393 000	18
AON	P. G. Ryan	1 384 000	17
St. Paul	D. W. Leatherdale	1 294 000	16
CAN Financial	D. H. Chookaszian	1 242 000	15
Providian	I. W. Bailey II	1 190 000	14
Jefferson-Pilot	D. A. Stonecipher	1 119 000	13

[16] Принято располагать блочные диаграммы вертикально, как сделано на этом рисунке, хотя не будет ошибкой и горизонтальное расположение.

Окончание табл. 4.2.1

Фирма	Руководитель	Зарплата и премии в 1994 году, дол.	Ранг
Aetna Life & Casualty	R. E. Compton	1 075 000	12
First USA	J. C. Tolleson	1 040 000	11
Salomon	R. E. Denham	1 000 000	10 Нижний квартиль = $1 000 000
Golden West Financial	H. M. Sandler	901 000	9
Cincinnati Financial	R. B. Morgan	896 000	8
Allstate	W. E. Hedien	767 000	7
Block (H&R)	T. M. Bloch	746 000	6
Franklin Resources	C. B. Johnson	743 000	5
Safeco	R. H. Eigsti	601 000	4
Equifax	C. B. Rogers, Jr.	554 000	3
Unintrin	R. C. Vie	481 000	2
Berkshire Hathaway	W. E. Buffet	100 000	1 Наименьшее значение = $100 000

*Это выброс.
Данные взяты из "Executive Compensation Scoreboard", *Business Week*, 1995, April, 24, p. 106–108.

Обратите внимание, насколько информативнее верхний рисунок, содержащий помеченные исключительные значения выплат руководителям отдельных фирм, по сравнению с нижним рисунком, на котором показано только пять базовых показателей. Хотя выше всего оплачиваются руководители некоторых финансовых компаний (выбросы), в целом размеры выплат в этой отрасли не очень отличаются от выплат руководителям в банковской сфере и в фармацевтической отрасли. Из рисунка также видно, что руководители коммунальных служб, за некоторыми исключениями, оплачиваются ниже, чем в других отраслях. Неплохо работать в отрасли, где нижний квартиль выплат составляет один миллион долларов в год!

Какая из диаграмм лучше? Есть смысл тратить время и энергию на построение подробной блочной диаграммы (с показом отдельных выбросов), только если это дает действительно необходимую дополнительную информацию. Стратегически разумно сначала быстро нанести на диаграмму пять базовых показателей, а затем уже решать, стоит ли тратить время и усилия на дополнительные подробности. Конечно, если построение диаграммы выполняется с помощью компьютера, всегда (или почти всегда) следует отдавать предпочтение подробной блочной диаграмме.

Функция кумулятивного распределения показывает перцентили

Функция кумулятивного распределения данных представляется в виде графика, который показывает перцентили путем установления соответствия между данными и процентами. Поскольку на вертикальной оси откладываются проценты от 0% до 100%, а на горизонтальной — сами перцентили (т.е. значения данных), то, используя этот график, можно легко находить либо значение перцентиля при заданном значении процента, либо значение процента, соответствующее определенному значению данных.

Функция кумулятивного распределения состоит из вертикальных скачков высотой $1/n$ для каждого из n значений данных и горизонтальных отрезков,

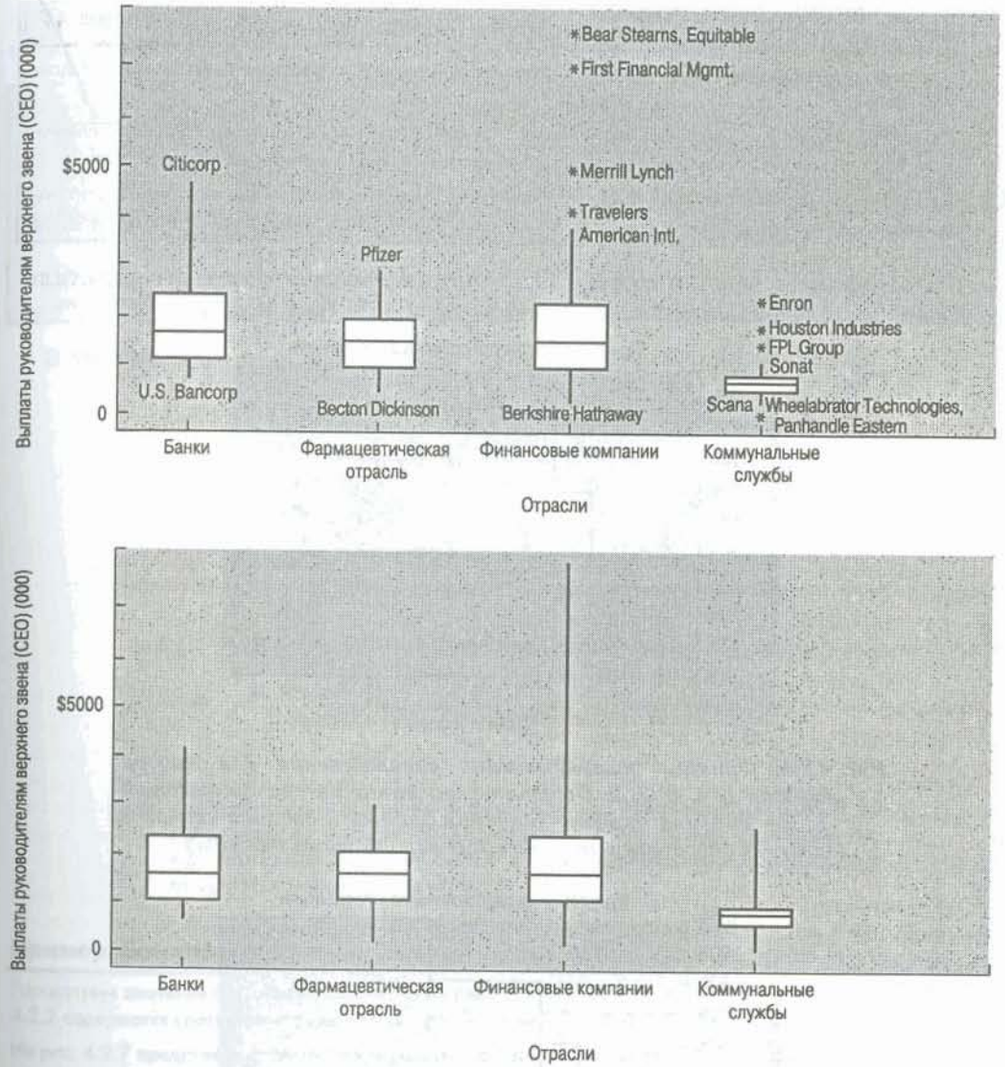


Рис. 4.2.3. Построенные в одном масштабе блочные диаграммы размеров выплат руководителям крупных фирм из некоторых отраслей позволяют легко сравнивать отрасли между собой. Верхний рисунок дополнительно содержит также выбросы и экстремумы, а на нижнем изображены только пять базовых показателей

соединяющих точки значений данных. На рис. 4.2.4 показана функция кумулятивного распределения для небольшого набора данных, состоящего из $n = 5$ значений (1, 4, 3, 7, 3), одно из которых (3) встречается дважды.

Если задано значение и необходимо найти его перцентильный ранг, необходимо поступать следующим образом.

Нахождение перцентильного ранга для заданного значения

1. Двигаясь по горизонтальной оси графика функции кумулятивного распределения, найдите заданное значение.
2. Двигайтесь вертикально вверх до пересечения с графиком функции кумулятивного распределения. Если вы попали на вертикальный участок, то переместитесь вверх на его середину.
3. Двигайтесь по горизонтали влево до пересечения с вертикальной осью, и вы получите перцентильный ранг.

В этом примере числу 4 соответствует 70-й перцентиль, так как перцентильный ранг этого значения расположен между 60 и 80% (рис. 4.2.5).

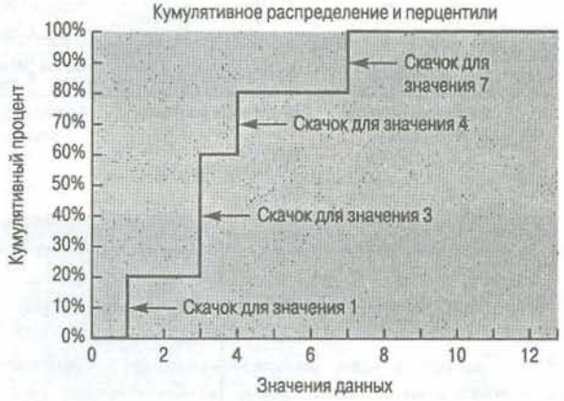


Рис. 4.2.4. Функция кумулятивного распределения для набора данных 1, 4, 3, 7, 3. Обратите внимание на скачок высотой 1 / n = 20% для каждого значения данных и двойной скачок в точке 3 (поскольку это значение встречается дважды)

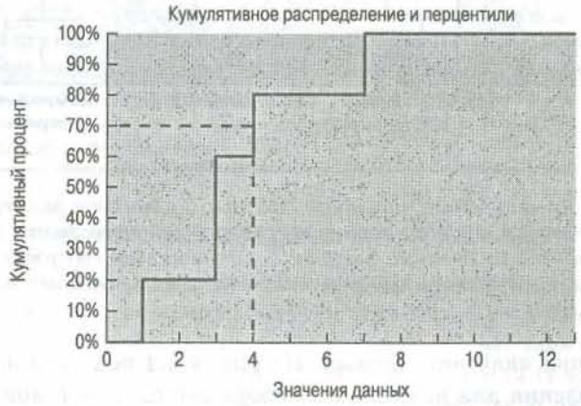


Рис. 4.2.5. Значение 4 представляет 70-й перцентиль. Двигайтесь вертикально вверх от значения 4, поскольку вы попали на вертикальный участок, переместитесь вверх на середину этого участка. Затем двигайтесь по горизонтали влево до пересечения с вертикальной осью, и вы получите результат 70%

Если задан процент, то соответствующий перцентиль можно найти следующим образом.

Нахождение перцентиля для заданного процента

1. Двигаясь по вертикальной оси графика функции кумулятивного распределения, найдите точку, соответствующую заданному проценту.
2. Двигайтесь вправо по горизонтали до пересечения с графиком функции кумулятивного распределения. Если вы попали на горизонтальный участок, то переместитесь к его середине.
3. От этой точки двигайтесь вертикально вниз. Точка пересечения с горизонтальной осью даст значение перцентиля.

В этом примере 44-му перцентилю соответствует число 3 (рис. 4.2.6).

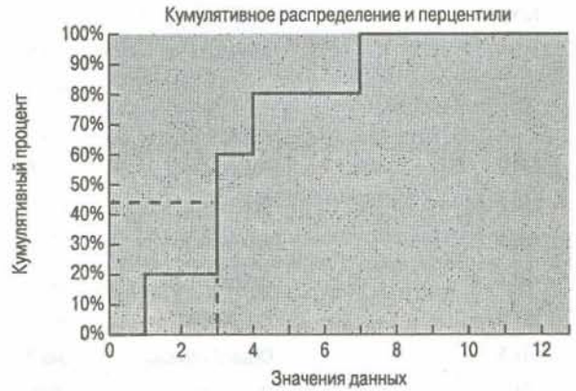


Рис. 4.2.6. Чтобы найти 44-й перцентиль, двигайтесь от 44% по горизонтали вправо до пересечения с графиком функции кумулятивного распределения и затем вертикально вниз, где получите искомое значение 3

Пример. Банкротства

Рассмотрим значения показателя количества банкротств на миллион человек в отдельных штатах. В табл. 4.2.2 содержатся соответствующие данные, упорядоченные по возрастанию.

На рис. 4.2.7 представлена функция кумулятивного распределения для этого набора данных. Из графика видно, что в большинстве штатов (от 10% до 90%) число банкротств находится в диапазоне от 150 до 400 банкротств на миллион населения.

На рис. 4.2.8 показано, как, используя функцию кумулятивного распределения, найти перцентили. Так, 50-й перцентиль равен 260,2 (штат Гавайи, как видно из данных таблицы), что соответствует значению медианы 260,2 банкротства на миллион человек. В то же время 90-й перцентиль равен 432,4 (штат Колорадо), а 95-й равен 524,4 (штат Аризона).

Для изображения данных вы можете выбрать любой из трех графиков: гистограмму, блочную диаграмму или график функции кумулятивного распределения. Все они отображают одну и ту же информацию (значения данных), но в различном виде. На рис. 4.2.9 приведены все три типа графического представления данных о количестве банкротств, что позволяет сравнить их между собой.

Областям высокой концентрации данных (т.е. тем, где находится большое количество значений) соответствуют пики на гистограмме и крутая функция кумулятивного распределения. Обычно, как и в нашем

случае, область высокой концентрации данных находится в середине. Областям низкой концентрации данных соответствуют низкие столбики на гистограмме и *пологий* участок кумулятивной кривой.

Блочная диаграмма содержит пять базовых показателей, которые можно увидеть и на функции кумулятивного распределения: наименьшее значение (для 0%), нижний квартиль (для 25%), медиана (для 50%), верхний квартиль (для 75%) и наибольшее значение (для 100%).

Таблица 4.2.2. Количество банкротств на миллион человек в отдельных штатах (данные упорядочены по возрастанию)

Штат	Количество банкротств	Штат	Количество банкротств
Арканзас	76,7	Виргиния	267,8
Южная Каролина	107,6	Мичиган	268,6
Миссисипи	121,8	Нью-Мексико	277,2
Луизиана	154,6	Вермонт	300,3
Северная Каролина	171,9	Мэн	309,1
Западная Виргиния	173,1	Мэриленд	310,2
Иллинойс	179,0	Айдахо	318,5
Айова	180,2	Орегон	319,6
Аляска	180,3	Коннектикут	333,5
Юта	188,7	Джорджия	339,7
Индиана	191,0	Род-Айленд	344,0
Вайоминг	191,5	Округ Колумбия	346,0
Огайо	191,8	Нью-Джерси	360,8
Делавэр	195,7	Флорида	372,0
Алабама	200,9	Нью-Йорк	380,1
Миннесота	203,9	Вашингтон	385,3
Монтана	206,2	Техас	393,5
Кентукки	222,0	Невада	408,9
Северная Дакота	228,3	Канзас	422,4
Миссури	235,0	Колорадо	432,4
Теннеси	237,1	Оклахома	445,7
Висконсин	243,0	Массачусетс	452,4
Южная Дакота	244,8	Аризона	524,4
Небраска	248,3	Нью-Гемпшир	548,4
Пенсильвания	259,3	Калифорния	631,0
Гавайи	260,2		

Данные вычислены из таблиц 26 и 847 Бюро переписи населения США *Statistical Abstract of the United States*, 1994, 114th ed. (Washington, D. C., 1994). Источник данных: Dun & Bradstreet Corporation, New York, NY, Business Failure Record, annual (авторское право защищено).

Обратите внимание, что из представленных здесь графических изображений данных только функция кумулятивного распределения содержит всю информацию о данных. При построении гистограммы часть информации теряется, так как гистограмма отражает только количество штатов в каждой из групп (например, *группа с количеством банкротств* от 100 до 200). При использовании блочной диаграммы также теряется часть информации, поскольку диаграмма содержит только пять базовых показателей. И лишь функции кумулятивного распределения содержат достаточно информации для того, чтобы можно было восстановить каждое число исходного набора данных.

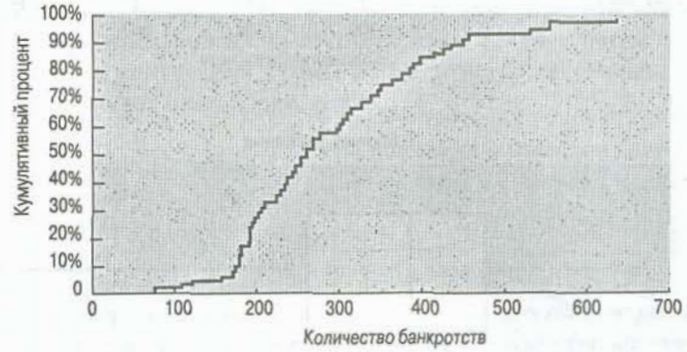


Рис. 4.2.7. Функция кумулятивного распределения для количества банкротств на миллион человек населения (по штатам) в 1993 г.

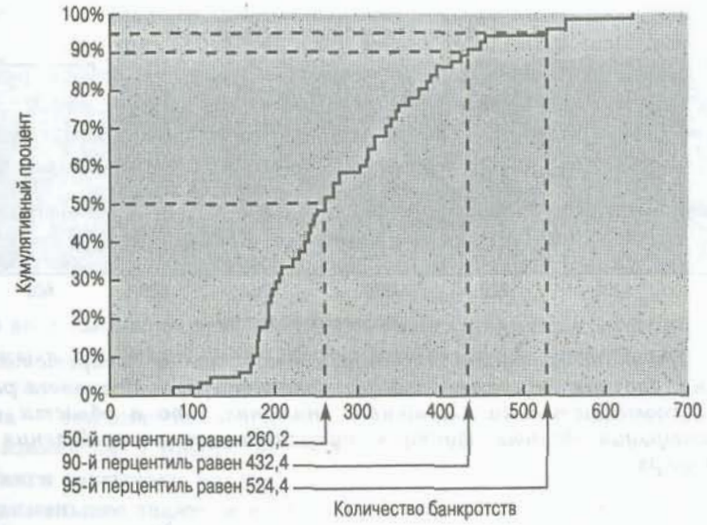


Рис. 4.2.8. Функция кумулятивного распределения для количества банкротств с отмеченными 50-м, 90-м и 95-м перцентилями

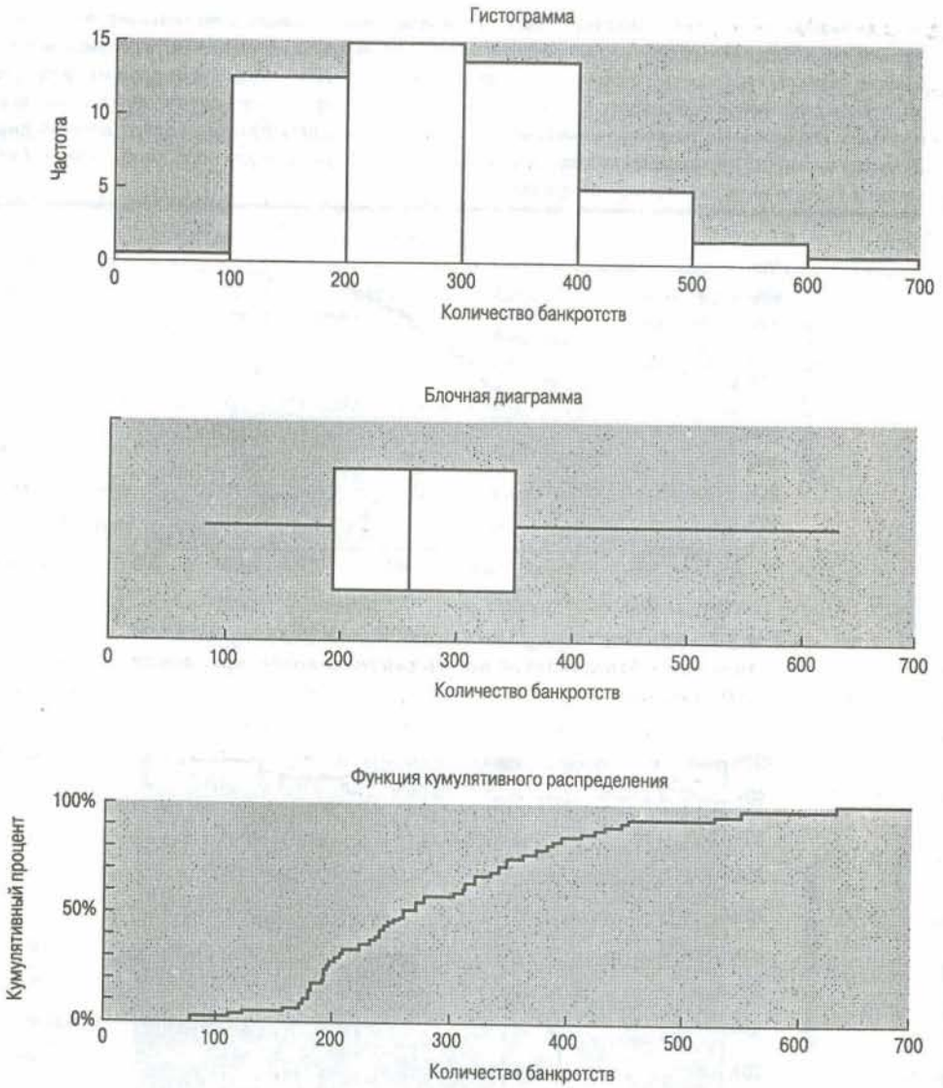


Рис. 4.2.9. Три типа графиков для данных о количестве банкротств: гистограмма, блочная диаграмма и график функции кумулятивного распределения соответственно. Обратите внимание, что в области высокой концентрации данных функция кумулятивного распределения круто идет вверх

4.3. Дополнительный материал

Резюме

Обобщение заключается в том, чтобы использовать один или несколько отобранных или рассчитанных значений для характеристики набора данных. При выполнении процедуры обобщения сначала следует описать основную структуру большинства значений данных, а затем все исключения или выбросы значений.

Среднее является наиболее часто используемым показателем типического значения в перечне значений данных. Вычисляют среднее путем сложения всех значений и деления полученной суммы на количество слагаемых. Формула вычисления среднего имеет следующий вид:

$$\text{Выборочное среднее} = \frac{\text{Сумма значений элементов данных}}{\text{Количество элементов данных}};$$

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Если речь идет о генеральной совокупности, то количество элементов принято обозначать N и значение среднего генеральной совокупности обозначать μ (греческая буква "мю"). Среднее распределяет общую сумму значений равномерно между всеми наблюдениями, и использовать его целесообразно тогда, когда в данных отсутствуют экстремальные значения (выбросы) и общая сумма значений важна для анализа. Среднее вычисляют только для количественных данных.

Взвешенное среднее (средневзвешенное) похоже на среднее, однако этот показатель позволяет присвоить каждому элементу данных свой "вес" (характеристику его важности). Это позволяет вычислять среднее в ситуациях, когда одни наблюдения более важны, чем другие, а значит, должны вносить больший вклад в результат. Формула вычисления взвешенного среднего (средневзвешенного) имеет следующий вид:

$$\text{Средневзвешенное} = \text{Сумма (вес умноженный на значение элемента)} =$$

$$= \omega_1 X_1 + \omega_2 X_2 + \ldots + \omega_n X_n = \sum_{i=1}^{n} \omega_i X_i .$$

Обычно веса выбирают так, чтобы их сумма была равна 1 (если это не так, то можно каждый вес разделить на общую сумму весов). Средневзвешенное можно вычислять только для количественных данных.

Медиана — это значение элемента, приходящееся на *середину* совокупности; половина элементов в наборе данных больше медианы, а вторая половина — меньше. **Ранги** связывают числа 1, 2, 3,... n со значениями данных таким образом, что наименьшее значение имеет ранг 1, следующее по величине значение — ранг 2 и т.д. до наибольшего значения, которое имеет ранг n. Ранг медианы $(1+n)/2$ показывает, сколько наблюдений следует отсчитать от наименьшего (или от наибольшего) значения, чтобы получить медиану. Если ранг медианы выражается не целым числом (например, 13,5 для $n = 26$), то усредняют два значения, расположенных по обе стороны от этого значения ранга (например, в

нашем случае — значения с рангами 13 и 14). Медиану можно вычислить как для количественных, так и для порядковых данных (упорядоченных категорий).

Медиана иначе, чем среднее, обобщает "типическое" значение. В то же время эти два значения близки между собой или совпадают, когда распределение симметрично (как, например, нормальное распределение). Если распределение асимметричное или содержит выбросы, то медиана и среднее могут различаться очень сильно.

Мода представляет собой наиболее распространенную категорию, т.е. такую, которая чаще других встречается в наборе данных. Моду можно вычислить для данных любого типа: количественных, порядковых и номинальных (неупорядоченных категорий). Для номинальных данных мода является единственной обобщающей характеристикой. Для количественных данных моду часто определяют как значение, соответствующее середине наивысшего столбика на гистограмме. Однако такое определение не совсем однозначно, поскольку середина столбика может зависеть от масштаба, в котором построена гистограмма.

Выбор обобщающей характеристики для конкретного набора данных необходимо осуществлять следующим образом. Для *номинальных* данных можно использовать только моду. Для *порядковых* данных можно использовать и моду, и медиану; мода выражает наиболее часто встречающуюся категорию, а медиана указывает категорию, расположенную в центре упорядоченного ряда значений. Для *количественных* данных можно использовать все три показателя. Если данные распределены приблизительно нормально, то значения всех трех показателей близки между собой и лучше всего использовать среднее. При асимметричном распределении эти три показателя могут существенно отличаться. В целом хороший результат дает медиана, поскольку она менее чувствительна к наличию экстремальных значений в области длинного хвоста кривой распределения. Однако если важна общая сумма значений, то предпочтительнее использовать среднее.

Перцентили выражают ранги как проценты от 0 до 100%, а не как числа от 1 до n; 0-й перцентиль соответствует наименьшему значению, 100-й перцентиль — наибольшему значению, 50-й — медиане и т.д. Отметим, что перцентиль измерен в тех же единицах, что и значения исходного набора данных (т.е. в долларах, галлонах и т.п.). Перцентили можно использовать для определения значения данных при заданном перцентильном ранге или, наоборот, для нахождения перцентильного ранга по заданному значению. Представляют также интерес **экстремумы** — *наибольшее* и *наименьшее* значения данных. **Квартили** — это 25-й и 75-й перцентили, ранги которых определяют по следующим формулам:

$$\text{Ранг нижнего квартиля} = \frac{1+\text{int}\left[(1+n)/2\right]}{2};$$

$$\text{Ранг верхнего квартиля} = n+1-\text{Ранг нижнего квартиля},$$

где int — функция взятия целого значения, которая отбрасывает дробную часть числа.

Пять базовых показателей набора данных включают наименьшее и наибольшее значения, нижний и верхний квартили и медиану. На **блочной диаграмме** эти пять показателей изображены в графической форме. **Выбросы** определяют как такие точки данных (если они есть), значения которых лежат далеко от тех значений, которые находятся в средней части набора данных. **Подробная блочная диаграмма** содержит значения выбросов с соответствующими метками, а также наиболее экстремальные из тех наблюдений, которые не являются выбросами. Для сравнения нескольких наборов данных, измеренных в одинаковых единицах, можно, используя один масштаб для каждого из них, построить блочную диаграмму и расположить эти диаграммы на одном рисунке.

Функция кумулятивного распределения данных представляется в виде графика, который показывает перцентили путем установления соответствия между данными и процентами. Этот график имеет вертикальный скачок величиной $1/n$ для каждого из n значений данных. Зная процент, можно найти перцентиль, двигаясь по графику вправо, а затем вниз. Зная значение, можно определить перцентильный ранг (процент), двигаясь по графику вверх и затем влево. Таким образом, функция кумулятивного распределения отражает перцентили и позволяет их вычислить. Это единственная графическая форма представления данных, которая "архивирует" данные, сохраняя достаточно информации для восстановления всех значений набора данных. Функция кумулятивного распределения круто растет в областях высокой концентрации данных (там, где высокие столбики на гистограмме).

Основные термины

- Обобщение (summarization), 117
- Усреднение (average), 118
- Среднее (mean), 118
- Взвешенное среднее (weighted average), 121
- Медиана (median), 126
- Ранг (rank), 126
- Мода (mode), 133
- Перцентиль (percentile), 137
- Экстремумы (extremes), 137
- Квартили (quartiles), 137
- Пять базовых показателей (five-number summary), 138
- Блочная диаграмма (box plot), 139
- Подробная блочная диаграмма (detailed box plot), 139
- Выброс (outlier), 139
- Функция кумулятивного распределения (cumulative distribution function), 142

Контрольные вопросы

1. Что представляет собой процесс обобщения набора данных? Почему так важно обобщать данные?
2. Перечислите и кратко опишите различные показатели, обобщающие данные.
3. Что необходимо делать с исключениями при обобщении набора данных?
4. Что означает типическое значение для перечня чисел? Назовите три различных способа определения типического значения.
5. Что такое среднее? Объясните среднее с точки зрения суммы всех значений набора данных.
6. Что такое взвешенное среднее? В каких случаях этот показатель используют вместо обычного среднего?
7. Что такое медиана? Как можно найти медиану исходя из ее ранга?
8. Как найти медиану для набора данных:
   а) С четным количеством значений?
   б) С нечетным количеством значений?
9. Что такое мода?
10. Как обычно определяют моду для количественного набора данных? Почему такое определение содержит неоднозначность?
11. Какой обобщающий показатель (или показатели) можно использовать для:
    а) Номинальных данных?
    б) Порядковых данных?
    в) Количественных данных?
12. Какие показатели лучше использовать при:
    а) Нормальном распределении данных?
    б) При планировании общего количества (суммы)?
    в) При асимметричном распределении, когда общая сумма не важна?
13. Что такое перцентиль? В частности, является ли он процентом (например, 23%) или выражен в тех же единицах, что и данные (например, $ 35,62)?
14. Назовите два способа использования перцентилей.
15. Что такое квартили?
16. Назовите пять базовых характеристик распределения.
17. Что такое блочная диаграмма? Какие детали часто дополнительно изображают на блочной диаграмме?
18. Что такое выброс (сильно отклоняющееся значение)? Как можно решить, является ли данная точка выбросом?
19. Рассмотрите функцию кумулятивного распределения:
    а) Что она собой представляет?
    б) Как ее начертить?
    в) Для чего ее используют?
    г) Сравните ее с гистограммой и блочной диаграммой.

Задачи

1. Качество дневной продукции автомобильного предприятия измеряется количеством автомобилей, требующих доработок после сборки. Рассмотрим данные о качестве автомобилей за 15 дней:

   30, 34, 9, 14, 28, 9, 23, 0, 5, 23, 25, 7, 0, 3, 24

   а) Определите среднее дневного выпуска бракованных автомобилей.
   б) Определите медиану дневного выпуска бракованных автомобилей.
   в) Начертите гистограмму для этих данных.
   г) Определите значение моды дневного выпуска бракованных автомобилей для гистограммы, построенной в п. "в".
   д) Определите квартили.
   е) Определите экстремумы (наименьшее и наибольшее значения).
   ж) Начертите блочную диаграмму для данных.
   з) Начертите функцию кумулятивного распределения данных.
   и) Найдите 90-й перцентиль для набора данных.
   к) Определите перцентильный ранг для завтрашнего количества бракованных автомобилей, равного 29.
2. В табл. 4.3.1 содержатся суммы, потраченные в прошлом месяце вашими постоянными потребителями на покупку вашей продукции.
   а) Определите средний объем продаж на одного постоянного потребителя.
   б) Определите медиану и квартили.
   в) Начертите блочную диаграмму.
   г) Найдите выбросы значений, если таковые имеются.
   д) Начертите подробную блочную диаграмму.
   е) Кратко прокомментируйте различия между этими двумя блочными диаграммами.
   ж) Если бы вы могли расширить этот список постоянных потребителей и включить в него еще три и если бы характеристика покупок этих трех фирм была такой же, как и у остальных, какой месячный объем продаж можно было бы ожидать для этих 13 потребителей?
   з) Напишите резюме о том, что вы узнали об этих потребителях, используя статистические методы.

Таблица 4.3.1 Объемы продаж в прошлом месяце

Потребитель	Продажи, тыс. дол.	Потребитель	Продажи, тыс. дол.
Consolidated, Inc.	142	Associated, Inc.	93
International, Ltd.	23	Structural, Inc.	17
Business Corp.	41	Communications Co	174
Computer Corp.	10	Technologies, Inc.	420
Information Corp.	7	Complexity, Ltd.	13

3. Многие страны (но не США) имеют "налог на добавленную стоимость" (НДС), который выплачивают предприниматели исходя из того, какую стоимость они добавили к своей продукции (например, разность между выручкой от продаж и стоимостью материалов). Этот налог отличается от налога на продажи, поскольку при оплате покупки потребитель не видит, что он платит дополнительный налог. В табл. 4.3.2 содержится выраженная в процентах величина НДС в различных странах.

   а) Постройте гистограмму для этого набора данных и кратко опишите форму распределения.

   б) Определите средний размер НДС по всем странам.

   в) Найдите медиану значения НДС.

   г) Сравните среднее и медиану. Соответствует ли это различие тому, что вы могли ожидать для такой формы распределения?

   д) Начертите функцию кумулятивного распределения.

   е) Какое значение НДС отвечает 20-му перцентилю? А какое 80-му перцентилю?

   ж) Найдите перцентиль, соответствующий размеру НДС 10%.

4. Рассмотрите прибыли крупных торговых фирм за 1997 г., представленные в табл. 4.3.3.

   а) Постройте для этого набора данных гистограмму и кратко опишите форму распределения.

   б) Найдите среднюю прибыль.

   в) Определите медиану уровня прибыли.

   г) Сравните среднее и медиану; в частности, какой из этих двух показателей больше? Ожидали ли вы такое различие при данной форме распределения?

Таблица 4.3.2. Величина НДС в различных странах

Страна	Величина НДС, %	Страна	Величина НДС, %
Бельгия	19,5	Голландия	18,5
Канада	7,0	Новая Зеландия	12,5
Дания	15,0	Норвегия	22,0
Франция	18,6	Португалия	16,0
Германия	15,0	Испания	15,0
Греция	18,0	Швейцария	6,5
Италия	19,0	Турция	12,5
Япония	3,0	Великобритания	17,5
Люксембург	15,0		

Данные взяты из Gilbert E. Metcalf, "Value-Added Taxation: A Tax Whose Time Has Come?". *Journal of Economic Perspective* 9, No. 1, (Winter 1995), p. 129. Источник: "Price Waterhouse Guide to Doing Business in..." для разных стран, Eurostat (1993), и OECD Revenue Statistics.

   д) Начертите функцию кумулятивного распределения.

   е) Ваша фирма имеет стратегический план увеличения прибыли до $50 000000. Какой перцентиль представляет такое значение прибыли в этом наборе данных?

   ж) Стратегический план вашей фирмы указывает, что уровень прибыли в настоящее время должен достичь 60-го перцентиля. Какое значение прибыли представляет этот перцентиль в этом наборе данных?

5. Коэффициент "бета" (β) акций фирмы указывает степень, в которой изменения цены акций отслеживают изменения на фондовом рынке в целом. Этот коэффициент интерпретируют как рыночный риск портфеля ценных бумаг. Коэффициент, равный 1, указывает на то, что в среднем стоимость акций

Таблица 4.3.3. Прибыли (1997 г.) крупных торговых фирм из Fortune 500

Фирма	Прибыль, млн дол.
Wal-Mart Stores	3526
Sears Roebuck	1118
Kmart	249
J. C. Penney	566
Dayton Hudson	751
Federated Department Stores	536
May Department Stores	775
Dillard's	258
Nordstrom	186
Harcourt General	– 115
PROFFITT'S	63
Mercantile Stores	130
Kohl's	141
Dollar General	145
Caldor	–155
Shopko Stores	49
Ames Department Stores	35
Saks Holdings	344
Family Dollar Stores	75
Fingerhut	69
Venture Stores	–195
Bradlees	–108
Value City	4

Взято из Fortune 500 по адресу `http://www.pathfinder.com/fortune/fortune500/ind149.html` от 14 ноября 1998 года.

возрастет (или упадет) на тот же самый процент, что и фондовый рынок в целом. Коэффициент, равный 2, указывает, что стоимость акций возрастает (или падает) в два раза больше по сравнению с фондовым рынком. Коэффициент "бета" портфеля ценных бумаг является средневзвешенным коэффициентом "бета" отдельных ценных бумаг, веса которым присвоены в зависимости от текущей рыночной стоимости ценных бумаг (рыночная стоимость вычисляется как произведение стоимости одной акции на количество акций). Рассмотрим следующий портфель ценных бумаг:

- 100 акций компании Speculative Computer по \$35 за акцию, $\beta = 2{,}4$;
- 200 акций компании Conservative Industries по \$88 за акцию, $\beta = 0{,}6$;
- 150 акций компании Dependable Conglomerate по \$53 за акцию, $\beta = 1{,}2$.

а) Определите значение "бета" для этого портфеля ценных бумаг.

б) Для снижения риска портфеля ценных бумаг вы решили продать все акции Speculative Computer и использовать полученные средства для покупки как можно большего количества акций Dependable Conglomerate17. Опишите новый портфель ценных бумаг, найдите его значение "бета" и убедитесь в снижении рыночного риска.

6. Ваша фирма имеет следующие ценные бумаги: обычные акции (рыночная цена \$4 500 000; инвесторы требуют 17% годовой нормы прибыли), привилегированные акции (рыночная стоимость \$1 700 000; текущая годовая доход составляет 13%) и 20-летние облигации (рыночная стоимость \$2 200 000; текущая годовая доходность составляет 11%). Определите стоимость вашего капитала.

7. Активные потребители составляют 13,6% рынка и тратят в среднем в \$16,23 в месяц на покупку вашей продукции. Пассивные потребители составляют 23,8% рынка и тратят \$9,85. Остальные потребители в среднем тратят \$14,77. Найдите средний объем трат для всех потребителей.

8. Опрос 613 человек, проживающих в регионе деятельности вашей фирмы, показал, что они в сумме собираются израсходовать на вашу продукцию в следующем году \$2135. Вы собираетесь расширить сферу деятельности вашей фирмы на город с населением 2,1 миллиона жителей.

   а) Определите среднюю сумму, которую потратит один человек, основываясь на данных опроса в том регионе, где сейчас работает ваша фирма.

   б) Какой годовой уровень продаж вы ожидаете при условии, что ваше присутствие на рынке нового города будет таким же, как и в том регионе, где вы уже работаете?

   в) Каким будет годовой уровень продаж, если вы ожидаете, что ваше присутствие на рынке нового города будет составлять только 60% по сравнению с регионом, где вы уже работаете?

[17] Будем считать, что есть возможность покупать и продавать любое количество акций. Не учитывайте реальные проблемы "неполных лотов" ("odd lots"), размер которых меньше, чем 100 акций.

9. Ваш отдел маркетинга выделяет четыре группы людей (типы A, B, C и D, где D означает "все остальные") в соответствии с индивидуальными чертами характера. Вы считаете, что в течение двух лет 38% представителей группы A купят ваше новое изделие. Аналогично, для группы B — 23%, для группы C — 8% и для группы D — 3%. Предположим, что в вашей целевой совокупности эти группы составляют 18, 46, 25 и 11% соответственно. Какую сумму продаж следует ожидать?

10. Ваш большой развлекательный комплекс под открытым небом имеет три въезда. По данным автоматического счетчика в прошлом году через первый вход въехало 11 967 транспортных средств, через второй — 24 205 и через третий — 7474. Проведенное исследование показало, что средняя запланированная длительность пребывания транспортных средств, проехавших через первый вход, составила 3,5 дня, въехавших через второй — 1,3 дня и въехавших через третий — 6 дней. Оцените типическое для всего комплекса в целом значение запланированной длительности пребывания одного транспортного средства

11. В табл. 4.3.4 приведены данные о размере местного налога и о населении северо-восточных центральных штатов. Определите размер местного налога на одного человека для всего данного региона[18].

12. Вы начали кампанию по улучшению качества продукции на вашей бумажной фабрике и для этих целей собрали большое количество докладных записок о проблемах потребителей. Представленная в каждой докладной записке проблема кодируется следующим образом: A — отсутствие бумаги; B — бумага слишком толстая; C — бумага слишком тонкая; D — ширина бумаги не соответствует стандарту; E — не тот цвет бумаги; F — края бумаги грубо обрезаны. Собранная информация приведена ниже:

    A, A, E, A, A, A, B, A, A, A, B, A, B, F, F, A, A, A, A, A, B, A, A, A, A, C, D, F, A, A, E, A, C, A, A, A, F, F

    а) Обобщите этот набор данных, вычислив процент проблем каждого вида в общем количестве проблем.

    б) Обобщите эти данные, вычислив моду.

    в) Напишите краткую (один абзац) докладную записку для руководства с рекомендациями наиболее эффективных действий по улучшению сложившейся ситуации.

    г) Можно ли в этом случае вычислить среднее или медиану? Если да, то почему, и если нет, то почему?

13. Рассмотрим данные о размере платы за ссуду под залог дома, представленные в табл. 4.3.5. Плата за ссуду указана как процент от величины ссуды и представляет собой одноразовый платеж при возвращении ссуды.

    а) Найдите среднее значение платы за ссуду.

[18] Информация о численности населения и размере местных налогов в 1996 г. взята из данных *Statistical Abstract of the United States: 1997* (117th ed.), Washington, D. C., 1997, p. 28, 314.

Таблица 4.3.4. Население штатов и размер налога

Штат	Население, тыс. чел.	Налоги штата (на одного человека), дол.
Огайо	11 173	1401
Индиана	5841	1444
Иллинойс	11 847	1458
Мичиган	9594	1994
Висконсин	5160	1864

Таблица 4.3.5. Плата за ссуду под залог дома

Фирма	Плата за ссуду, %	Фирма	Плата за ссуду, %
Allied Pacific Mortgage	1,25	Mortgage Associates	2
Alternative Mortgage	2	Normandy Mortgage	1,25
Bankplus Mortgage	1	Performance Mortgage	2
Bay Mortgage	2	PNC Mortgage	1
CTX Mortgage	1,5	Qpoint Home Mortgage	1,5
First Mark Mortgage	2	Sammamish Mortgage	2
Mariner Mortgage	1	U. S. Discount Mortgage	2

Взято из "Summer Mortgage Rates". *The Seattle Times*, 1995, July, 16, p. G1.

б) Определите медиану платы за ссуду.

в) Найдите моду.

г) Какой из показателей (среднее, медиана или мода) наиболее полезен для описания типического значения платы за ссуду? Почему?

14. Торгующая по почте компания первоначально разослала свой новый каталог репрезентативной выборке из 10000 человек, взятых из имеющегося списка рассылок, и получила заказы на общую сумму $36851.

а) Определите средний размер заказа (в долларах) на одного человека из этой первоначальной рассылки.

б) Какой общий объем заказов (в долларах) следует ожидать компании в случае рассылки каталога всем 563000 клиентам, включенным в список рассылки.

в) В выборке из 10000 человек, объем заказов для которой составил $36851, реально только 973 человека действительно сделали заказ. Определите средний объем заказа для тех, кто действительно сделал заказ.

г) Исходя из п. "в" определите, сколько заказов следует ожидать компании после того, как каталог будет послан каждому из 563000 человек, включенных в список рассылки?

15. Рассмотрим прочность хлопковых нитей на ткацкой фабрике (в фунтах силы на разрыв) для выборки нитей, взятых со склада:

117, 135, 94, 79, 90, 85, 173, 102, 78, 85, 100, 205, 93, 93, 177, 148, 107.

а) Определите среднее значение прочности нити на разрыв.

б) Определите медиану прочности нити на разрыв.

в) Постройте гистограмму, отметьте среднее и медиану, кратко прокомментируйте отношение между ними. Одинаковы ли значения этих двух показателей? Если да, то почему, и если нет, то почему?

г) Начертите кумулятивное распределение.

д) Найдите 10-й и 90-й перцентили.

е) Руководство фабрики хотело бы получать от поставщиков по крайней мере 90% нитей с прочностью на разрыв 100 фунтов или больше. Исходя из приведенного набора данных решите, соответствует ли поставленное сырье такому качеству? В частности, с каким перцентилем вы будете проводить сравнение?

16. За прошедший год уровень материальных запасов на вашей фабрике измеряли 12 раз, результаты приведены ниже. Определите средний уровень материальных запасов в течение года:

313, 891, 153, 387, 584, 162, 742, 684, 277, 271, 285, 845.

17. Следующий список представляет долю вашей продукции в 20 главных регионах страны:

0,7%; 20,8%; 2,3%; 7,7%; 5,6%; 4,2%; 0,8%; 8,4%; 5,2%; 17,2%; 2,7%; 1,4%; 1,7%; 26,7%; 4,6%; 15,6%; 2,8%; 21,6%; 13,3%; 0,5%.

а) Определите среднее и медиану.

б) Начертите функцию кумулятивного распределения для этого набора данных.

в) Определите 80-й перцентиль.

18. Рассмотрите месячные объемы продаж (в тысячах долларов) 17 торговых представителей вашей фирмы:

23, 14, 26, 22, 28, 21, 34, 25, 32, 32, 24, 34, 22, 25, 22, 17, 20.

а) Найдите среднее и медиану.

б) Постройте блочную диаграмму.

19. Рассмотрите процентное изменение курса доллара по отношению к другим валютам в течение четырех недель (табл. 4.3.6).

а) Определите среднее процентное изменение курса доллара, усреднив значения по всем представленным странам.

б) Как вы считаете, в течение рассматриваемого периода времени в среднем доллар усилился или ослаб по отношению к другим валютам?

в) Определите медиану. Почему в данном случае она настолько отличается от среднего?

г) Постройте блочную диаграмму.

Таблица 4.3.6. Изменение курса доллара

Страна	Изменение, %	Страна	Изменение, %
Бельгия	–5,3	Сингапур	–1,5
Япония	–6,7	Франция	–4,9
Бразилия	26,0	Южная Корея	–1,0
Мексика	–1,2	Гонконг	0,0
Великобритания	–3,7	Тайвань	–0,1
Голландия	–5,1	Италия	–4,7
Канада	–1,9	Германия	–5,1

20. Если бы вы имели список данных, отражающих количество миль на один галлон бензина для различных автомобилей, то как бы мог выглядеть 65-й перцентиль для таких данных: 65 автомобилей, 65%, $13860 или 27 миль на один галлон?

21. Для данных о доходности необлагаемых налогом облигаций (задача 6 из главы 3):

    а) Определите среднюю доходность.

    б) Определите медиану доходности.

    в) Определите квартили.

    г) Найдите пять базовых показателей.

    д) Начертите блочную диаграмму для этих значений доходности.

    е) Начертите функцию кумулятивного распределения для этого набора данных.

    ж) Найдите значение перцентиля для 5,90.

    з) Найдите величину 85-го перцентиля.

22. Используя данные задачи 7 из главы 3, вычислите среднее и медиану для определения типической реакции рынка на объявление фирмы о скупке собственных акций.

23. Используя данные задачи 9 из главы 3, для портфеля инвестиций компании College Retirement Equities Fund (CREF) в магазины мебели и интерьера:

    а) Определите среднюю рыночную стоимость акции каждой из фирм в портфеле ценных бумаг CREF.

    б) Определите медиану этих рыночных стоимостей.

    в) Сравните среднее с медианой.

    г) Найдите пять базовых показателей для этих данных.

    д) Постройте блочную диаграмму и прокомментируйте форму распределения. В частности, есть ли признаки асимметрии?

    е) Соответствует ли взаимосвязь между средним и медианой форме распределения? Если да, то почему, и если нет, то почему?

24. В табл. 4.3.7 содержатся данные о продолжительности подборки видеофильмов.

    а) Определите среднюю продолжительность фильмов.

    б) Определите медиану продолжительности фильмов.

    в) Что больше: среднее или медиана? Исходя из вашего ответа, будете ли вы ожидать сильную асимметрию в направлении больших значений?

    г) Начертите гистограмму и прокомментируйте ее связь с вашим ответом в п. "в".

25. На собраниях некоторой группы людей показывают только фильмы, длительность которых не превышает 100 минут. Рассмотрите приведенные в табл. 4.3.7 данные о продолжительности подборки видеофильмов.

    а) Какой процент этих фильмов можно показать данной группе?

    б) Назовите самый продолжительный из тех фильмов, которые можно показать данной группе.

    в) Прокомментируйте взаимосвязь между вашим ответом в п. "а" и перцентильным рангом вашего ответа в п. "б".

26. В винном магазине продается 86 наименований вина урожая 1992 года, 125 наименований вина урожая 1993 года, 73 наименований вина урожая 1994 года и 22 наименования вина урожая 1995 года. Будем рассматривать наименование (марку) вина как элементарную единицу анализа.

    а) Определите моду года урожая вина. О чем свидетельствует этот показатель?

    б) Определите среднее значение года урожая и сравните его с модой.

    в) Начертите гистограмму для года урожая.

    г) Если средняя цена продажи вина урожая 1992 года составляет $7,99, вина урожая 1993 года — $7,74, вина урожая 1994 года — $8,57, вина урожая 1995 года — $6,99, найдите среднюю цену продажи для всех этих вин. (*Подсказка*: будьте внимательны!)

Таблица 4.3.7. Продолжительность подборки видеофильмов

Продолжительность, мин.	Фильм	Продолжительность, мин.	Фильм
133	Flower Drum Song	84	Origins of American Animation
111	Woman of Paris, A	109	Dust in the Wind (Chinese)
88	Dim Sum: A Little Bit of Heart	57	Blood of Jesus, The
120	Do the Right Thing	60	Media: Zbig Rycznskia Collection
87	Modern Times	106	Life (Tape 2) (Chinese)
100	Law of Desire (Spanish)	101	Dodsworth
104	Crowd, The	123	Rickshaw Boy (Chinese)
112	Native Son	91	Gullivar's Travels
134	Red River	136	Henry V (Olivier)
99	Top Hat		

27. Вернитесь к примеру об изменениях в объеме затрат на телевизионную рекламу из главы 3 (табл. 3.6.1). Мы определили два выброса, но можно выделить еще два. Постройте подробную блочную диаграмму для этого набора данных, указав конкретно, какие рекламодатели представляют собой сильно отличающиеся значения (выбросы).

28. Рассмотрите данные о расходах на лечение заболеваний сердца в больницах района Puget Sound из задачи 12 главы 3.

    а) Обобщите размеры расходов на лечение.

    б) Начертите блочную диаграмму.

    в) Начертите функцию кумулятивного распределения.

    г) Если бы ваша больница захотела поместить себя в 65-й перцентиль по отношению к размерам расходов в этом регионе, то какой должна была бы быть цена лечения в вашей больнице?

29. Рассмотрите данные о размерах выплат главным должностным лицам (CEO) фирм, производящих продукты питания, из задачи 13 главы 3.

    а) Постройте подробную блочную диаграмму.

    б) Определите 10-й перцентиль размера выплат.

30. Исходя из данных задачи 16 главы 3 охарактеризуйте с помощью среднего и медианы стоимость ритуальных услуг.

31. Используйте набор данных о низком качестве электромоторов из задачи 19 главы 3.

    а) С помощью среднего и медианы охарактеризуйте типический уровень бракованной продукции.

    б) Исключите два выброса и заново вычислите среднее и медиану.

    в) Сравните среднее и медиану, вычисленные с учетом выбросов и без них. В частности, насколько чувствительны два этих показателя к наличию выбросов?

32. Многие маркетологи считали, что потребители в основном должны предпочитать низкокалорийные продукты питания. Эти "облегченные" продукты питания получили распространение, но они продавались не в таких больших количествах (за некоторым исключением), как того хотелось бы их производителям. В табл. 4.3.8 содержатся данные об объемах продаж некоторых известных марок "облегченных" продуктов питания.

    а) Определите размер общего рынка продуктов этих марок.

    б) Определите среднее продаж для этих продуктов.

    в) Начертите функцию кумулятивного распределения.

    г) Ваша фирма планирует запустить в производство новый вид "облегченного" продута питания. Цель состоит в достижении, по крайне мере, 20-го перцентиля имеющихся видов. Определите целевой годовой объем продаж в долларах.

Таблица 4.3.8. Объемы продаж некоторых "облегченных" продуктов питания

"Облегченные" продукты питания	Объем продаж, млн дол.
Entenmann's Fat Free baked goods	125,5
Healthy Request soup	123,0
Kraft Free processed cheese	83,4
Aunt Jemima Lite and Butter Lite pancake syrup	58,0
Fat Free Fig Newtons	44,4
Hellmann's Light mayonnaise	38,0
Louis Rich turkey bacon	32,1
Kradt Miracle Whip free	30,3
Ben & Jerry's frozen yogurt	24,4
Hostess Lights snack cakes	19,3
Perdue chicken/turkey franks	3,8
Milky Way II candy bar	1,1

Данные взяты из "Light" Foods are Having Heavy Going". *The Wall Street Journal*, 1993, March, 4, p. B1. Источник: Information Resources Inc.

33. Для процентного изменения индекса Dow Jones Transportation Average по сравнению с 31 августа 1998 года (задача 24 из главы 2):

    а) Определите среднее процентного изменения.

    б) Определите медиану процентного изменения.

    в) Определите пять базовых показателей для процентного изменения.

    г) Постройте блочную диаграмму для процентного изменения.

    д) Начертите функцию кумулятивного распределения для процентного изменения.

    е) Определите перцентиль значения 10% и значение 90-го перцентиля.

34. Для значения индекса Dow Jones Transportation Average в сентябре 1998 года (задача 25 из главы 2):

    а) Определите среднее чистого изменения.

    б) Определите медиану чистого изменения.

    в) Определите пять базовых показателей для чистого изменения.

    г) Постройте блочную диаграмму для чистого изменения.

    д) Определите среднее процентного изменения.

    е) Определите медиану процентного изменения.

    ж) Определите пять базовых показателей для процентного изменения.

    з) Постройте блочную диаграмму для процентного изменения.

Упражнения с использованием базы данных

Обратимся к базе данных о служащих, приведенной в приложении А.

1. Для размеров годовой заработной платы:
   а) Определите среднее.
   б) Определите медиану.
   в) Постройте гистограмму и определите приблизительное значение моды.
   г) Сравните эти три показателя. Что вы можете сказать о типическом размере заработной платы в этом административном подразделении?
2. Для размеров годовой заработной платы:
   а) Начертите функцию кумулятивного распределения.
   б) Найдите медиану, квартили и экстремумы.
   в) Постройте блочную диаграмму и прокомментируйте ее.
   г) Определите 10-й и 90-й перцентили.
   д) Чему равен перцентильный ранг для служащего под номером 6?
3. Рассматривая пол служащих:
   а) Обобщите данные, вычислив процент мужчин и женщин.
   б) Найдите моду. О чем она свидетельствует?
4. В отношении возраста: ответьте на вопросы упр. 1.
5. В отношении возраста: ответьте на вопросы упр. 2.
6. В отношении стажа работы: ответьте на вопросы упр. 1.
7. В отношении стажа работы: ответьте на вопросы упр. 2.
8. В отношении уровня подготовки: ответьте на вопросы упр. 3.

Проекты

1. Используя Internet или экономические журналы из вашей библиотеки, подберите набор данных из 25 чисел, характеризующих интересующую вас фирму или отрасль промышленности. Обобщите эти данные, используя все изученные вами методы, которые можно применить в данном случае. Используйте как числовые, так и графические методы. Представьте результаты в виде краткой (две страницы) докладной записки, указав в первом абзаце свои рекомендации. (Не используйте большие графики.)
2. Найдите статистические характеристики для двух выбранных вами одномерных количественных наборов данных, которые связаны с вашей работой, фирмой или отраслью промышленности. Для каждого набора данных:
   а) Определите среднее, медиану и моду.
   б) Как каждый из этих показателей характеризует набор данных и экономическую ситуацию?

   в) Постройте гистограмму и укажите значения этих трех характеристик на горизонтальной оси. Прокомментируйте форму распределения и взаимосвязь между гистограммой и этими характеристиками.
   г) Постройте блочную диаграмму и прокомментируйте преимущества и недостатки гистограммы по сравнению с блочной диаграммой.

Ситуация для анализа

Управленческие прогнозы о производстве и маркетинге, или "Случай подозрительного потребителя"

Придя на работу, мистер Б. Р. Харрис, как и ожидал, обнаружил у себя на столе рекомендации мистера Х. Е. Макроури. В них содержались основные данные для квартальной презентации Харриса относительно объемов производства на следующие три месяца, которую он должен был провести сегодня для высшего руководства. Эти прогнозы должны были лечь в основу планирования и показать теоретически соответствующие объемы закупок, запасов и рабочих ресурсов в ближайшем будущем. Однако обычно потребители ведут себя вопреки ожиданию, поэтому подобные прогнозы всегда сложны и, как правило, включают элемент предположений (субъективного мнения).

Харрис и Макроури решили изменить традицию и подготовить более объективное обоснование для этих прогнозов. Макроури в последнее время анализировал данные опроса потребителей (новая экспериментальная процедура, основанная на ответах 30 репрезентативных потребителей) и подготовил отчет, в котором, в частности, утверждалось:

"В следующем квартале мы ожидаем объем продаж на сумму $1 477 108. Прогнозы объемов продаж по регионам приведены в таблице. Мы рекомендуем увеличить производство до уровня, который согласуется с ожидаемым ростом объемов продаж..."

	II квартал 1999 г. (прогноз), дол.	I квартал 1999 г. (факт), дол.	II квартал 1998 г. (факт), дол.
Объемы продаж:			
Северо-восток	441 058	331 309	306 718
Северо-запад	291 948	22 185	200 201
Юг	149 518	118 151	101 721
Средний запад	370 577	277 952	254 315
Юго-запад	224 007	165 332	157 843
Итого	1 477 108	1 114 929	1 020 798
Продукция (оптовая стоимость):			
Стулья	514 468	425 925	389 115
Столы	228 314	201 125	197 250

Окончание таблицы

	II квартал 1999 г. (прогноз), дол.	I квартал 1999 г. (факт), дол.	II квартал 1998 г. (факт), дол.
Книжные полки	272 624	209 105	189 475
Шкафы	461 702	276 500	295 400
Итого	1 477 108	1 112 655	1 071 240
Производство (количество штук):			
Стулья	11 433	9465	8647
Столы	1827	1609	1578
Книжные полки	4194	3217	2915
Шкафы	1319	790	844

Харрису было нелегко. Прогноз содержал большое увеличение объемов как по отношению к текущему кварталу (на 32,5%), так и к аналогичному кварталу прошлого года (на 44,7%). За последние годы темпы роста фирмы не были такими высокими. Вместе с тем, рекомендации содержали предложения об увеличении объема производства в связи с ожидаемым увеличением продаж. Почему возникают сомнения? Потому что, если прогноз неверный и объем продаж не увеличится, фирма получит большой и дорогостоящий запас готовой продукции (которая, к тому же, произведена с повышенными, по сравнению с обычными, затратами из-за оплаты сверхурочных работ, найма дополнительных рабочих и аренды дополнительного оборудования) в дополнение к своим обычным текущим затратам (включая процент, который фирма могла бы получить с суммы денег, которую она вынуждена была потратить на производство дополнительной продукции).

Харрис высказал свои сомнения, и Макроури тоже заколебался. Да, все казалось просто: взять из результатов опроса среднее прогнозируемое значение потребительских расходов и умножить его на общую численность потребителей в данном регионе. В чем может быть ошибка? Харрис и Макроури решили внимательнее изучить данные. Ниже приведена таблица, которая включает общую информацию (оптовая цена каждого наименования продукции и количество реальных покупателей по регионам) и результаты выборочного исследования. Каждый из 30 отобранных потребителей указал, сколько единиц каждого из наименований товара он планирует заказать в следующем квартале. Колонка "Стоимость" содержит объем наличных денег, которые получит фирма (например, покупатель 1 планирует приобрести 3 стула по $45 и 4 книжные полки по $65, на общую сумму $395).

Опт	Цена, дол.
Стулья	45
Столы	125
Книжные полки	65
Шкафы	350

Окончание таблицы

Опт	Цена, дол.
Реальные покупатели	
Северо-восток	303
Северо-запад	201
Юг	103
Средний запад	255
Юго-запад	154
Итого	1016

Результаты выборочного исследования

Покупатель (№)	Стулья, шт.	Столы, шт.	Книжные полки, шт.	Шкафы, шт.	Стоимость, дол.
1	3	0	4	0	395
2	9	1	6	1	1270
3	23	2	1	2	2050
4	7	0	3	0	510
5	4	0	0	0	180
6	14	1	5	0	1080
7	6	0	5	0	595
8	14	1	0	0	755
9	1	5	17	3	2825
10	2	0	4	1	700
11	16	1	1	1	1260
12	4	0	4	0	440
13	6	0	4	1	880
14	2	1	8	2	1435
15	42	15	21	18	11 430
16	3	0	0	2	835
17	7	3	0	0	690
18	1	4	2	0	675
19	43	0	4	0	2195
20	6	2	4	2	1480
21	3	1	1	0	325
22	45	6	1	0	2840
23	0	2	7	1	1055
24	13	6	3	0	1053

Окончание таблицы

Покупатель (№)	Стулья, шт.	Столы, шт.	Книжные полки, шт.	Шкафы, шт.	Стоимость, дол.
25	19	0	2	2	1685
26	0	0	0	0	0
27	8	0	7	0	815
28	14	3	3	1	1550
29	6	0	1	2	1035
30	17	0	6	0	1155
Итого (для выборки)	338	54	124	39	43 670
Среднее	11,267	1,8	4,133	1,3	1 455 667
Средняя стоимость	$ 507	$ 225	$ 268 667	$ 455	$ 1 455 667
Общий прогноз (результат умножения на 1015 покупателей):					
Стоимость	$ 514 468	$ 228 314	$ 272 624	$ 461 702	$ 1 477 108
Количество единиц	11 433	1827	4194	1319	

Вопросы для обсуждения

1. Подходит ли в данном случае обычно используемый Харрисом и Макроури метод, основанный на среднем, или этот метод изначально неверен? Обоснуйте ваш ответ.
2. Внимательно изучите данные, используя статистические характеристики и графики. Какой можно сделать вывод?
3. Что бы вы порекомендовали сделать Харрису и Макроури для подготовки к сегодняшней презентации?

ГЛАВА 5

Изменчивость: изучение разнообразия

Одна из причин, по которым возникает необходимость в проведении статистического анализа, состоит в том, что данные изменчивы. Если бы данные не изменялись, то ответы на многие вопросы были бы просто очевидными и нам не нужно было бы обращаться к методам статистического анализа[1]. Ситуация, в которой присутствует изменчивость, часто содержит риск, поскольку даже использование всей доступной информации не позволяет точно предугадать, что же произойдет в будущем. Для адекватной работы с риском необходимо понимать его природу и уметь измерять изменчивость (часто также используют термин "вариация". — *Прим. ред.*), которая является его следствием. Приведем несколько ситуаций, в которых изменчивость имеет важное значение.



Случай первый. Рассмотрим изменчивость производительности труда работников. Совершенно очевидно, что эффективность работы отдела определяется общей производительностью труда всех его сотрудников. Однако любые усилия, направленные на повышение производительности труда, должны учитывать индивидуальные особенности работников. Например, некоторые программы повышения производительности труда могут быть ориентированы на всех работников, в то время как другие — уделять особое внимание самым "быстрым" или самым "медленным" из них. Определение изменчивости производительности труда дает возможность выявить разброс таких индивидуальных различий и получить полезную информацию для планирования мероприятий по повышению общей производительности труда.

[1] Некоторые специалисты в области статистики в частных беседах отмечают, что именно наличие изменчивости данных обеспечивает им работу!

Случай второй. Фондовая биржа в среднем обеспечивает более высокую доходность вложенных средств, чем, например, фонды денежного рынка. Однако работа на фондовой бирже связана с большим риском, и инвестирование в акции может привести к реальным потерям. Таким образом, средняя, или "ожидаемая", доходность *не отражает полностью всю* картину. Мера изменчивости доходности отдельных инвестиций будет отражать уровень риска, сопряженного с каждым конкретным вложением средств.

Случай третий. Предположим, что вы сравниваете маркетинговые затраты своей фирмы с аналогичными затратами фирм, работающих в вашей отрасли промышленности, и обнаруживаете, что затраты вашей фирмы меньше затрат, типичных для данной отрасли. Для того чтобы оценить затраты на будущее, очень полезным может оказаться учет разброса соответствующих данных по отрасли. Найдя разность между значением затрат своей фирмы и средним значением по отрасли и сравнив полученную величину с мерой изменчивости затрат в отрасти, можно сделать вывод о том, находится ли маркетинговая деятельность вашей фирмы *в сравнении с другими аналогичными* фирмами лишь на несколько более низком уровне или же ваша фирма является некоторым исключением из общей картины. Такая информация может помочь в стратегическом планировании затрат на маркетинг в следующем году.

Изменчивость можно определить как степень различий между отдельными значениями. Подобный смысл имеют также такие понятия, как **разнообразие, неопределенность, рассеяние** и **разброс**. Далее мы увидим, что существуют три разных способа описания степени изменчивости набора данных, причем каждый из них требует соответствующих числовых значений.

1. *Стандартное отклонение* (в русскоязычной литературе по статистике часто также используют термины "среднее квадратическое отклонение" и "среднее квадратичное отклонение." — *Прим. ред.*) *используют* наиболее часто. Этот показатель описывает, насколько сильно результат наблюдений обычно отличается от среднего значения. При возведении стандартного отклонения в квадрат получаем *дисперсию*.
2. *Размах* легко вычисляется, однако дает несколько поверхностное представление об изменчивости данных и имеет ограниченное применение. Эта величина описывает пределы изменения данных в наборе и представляет собой расстояние между минимальным и максимальным значениями.
3. *Коэффициент вариации* обычно выбирается в качестве *относительной* (в противоположность *абсолютной*) меры изменчивости. Этот показатель используется достаточно часто. Он показывает, насколько сильно обычно отличается результат конкретного наблюдения от среднего значения, в процентном отношении к среднему; при этом используется отношение стандартного отклонения к среднему значению.

Мы также рассмотрим, как влияет на изменчивость данных изменение шкалы измерения (например, переход от японской иены к долларам США или переход от количества единиц выпущенной продукции к денежной стоимости этой продукции).

5.1. Стандартное отклонение: традиционный выбор

Стандартное отклонение — это число, описывающее, насколько значения данных обычно *отличаются от среднего*. Понятие стандартного отклонения является очень важным в статистике, поскольку оно представляет собой основной инструмент определения степени случайности в изучаемой ситуации. В частности, этот показатель является мерой случайности отклонений отдельных значений от их среднего.

Если все величины одинаковы, как, например, в приведенном ниже простом наборе данных

5,5; 5,5; 5,5; 5,5

то среднее будет иметь значение $\bar{X} = 5{,}5$, а стандартное отклонение составит $S = 0$. Последнее отражает тот факт, что в этом тривиальном наборе данные не подвержены изменчивости.

В реальной жизни большинство данных характеризуется большей или меньшей степенью изменчивости. Отдельные значения набора данных располагаются на некотором расстоянии от среднего, а стандартное отклонение характеризует степень изменчивости. Рассмотрим теперь другой набор данных, которым присуща некоторая изменчивость:

43,0; 17,7; 8,7; −47,4

Эти числа являются значениями ставки доходности (например, 43%) акций четырех компаний (*Maytag, Boston Scientific, Catalytica* и *Mitcham Industries*), выбранных случайным образом (случайность обеспечивалась путем метания стрелы игры дарт в газетную страницу с котировками акций)[2]. Среднее значение в этом случае такое же, $\bar{X} = 5{,}5$, т.е. эти акции имеют среднюю ставку доходности 5,5% (это означает, что портфель равных в денежном выражении инвестиций в названные выше акции будет иметь эту среднюю доходность 5,5%). Несмотря на то что среднее значение здесь такое же, как и в предыдущем случае, отдельные значения данных существенно различаются между собой. Первая величина, 43,0, располагается на расстоянии $X_1 - \bar{X} = 43{,}0 - 5{,}5 = 37{,}5$ от среднего значения. Из этого следует, что ставка доходности акций *Maytag* превышает среднюю ставку доходности на 37,5%. Последнее значение, −47,4, расположено от среднего на расстоянии $X_4 - \bar{X} = -47{,}4 - 5{,}5 = -52{,}9$; таким образом, ставка доходности акций *Mitcham Industries* оказывается на 52,9% *ниже* среднего уровня (ниже — поскольку величина отрицательна). В табл. 5.1.1 показано, насколько каждое из значений отличается от среднего.[3]

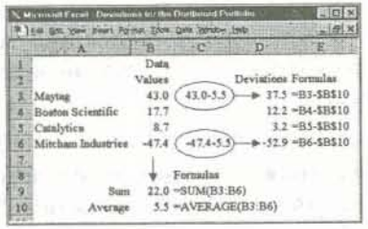
2 Georgette Jasen, "Your Money Matters: Winds of Chance Blow Cold on the Pros", *The Wall Street Journal*, April 9, 1998, p. C1.

3 Знак доллара в формулах дает Excel указание использовать одно и то же среднее значение (B10) в вычислениях со всеми последующими значениями данных. Это позволяет легко копировать формулу с переносом ее вниз по столбцу после ввода первый раз (в данном случае — в *ячейку D3*). Для того чтобы выполнить копирование, нужно выделить ячейку, а затем либо воспользоваться командами Edit⇨Copy (Правка⇨Копировать) и Edit⇨Paste (Правка⇨Вставить), либо просто перетащить нижний правый угол ячейки вниз по столбцу.

Описанные выше расстояния от среднего значения называются **отклонением**, или разностью. Они показывают, насколько выше среднего значения (в случае положительной разности) или ниже среднего (если разность отрицательна) лежит каждое значение данных. Отклонения в свою очередь образуют набор данных, расположенных вокруг нуля, что похоже на исходный набор данных, значения в котором расположены вокруг среднего.

В качестве обобщающей характеристики отклонений используют стандартное отклонение. Просто усреднить отклонения нельзя, поскольку часть из них окажется отрицательными, а часть — положительными, в результате чего результат такого усреднения всегда будет равен нулю и не будет содержать никакой дополнительной информации[4]. Вместо этого используют стандартный прием, заключающийся в том, что каждое значение сначала возводят в квадрат (т.е. его умножают на себя), чтобы избавиться от знака "минус", затем складывают, делят на n–1 и извлекают квадратный корень (это обратная операция по отношению к выполненному ранее возведению в квадрат)[5].

Таблица 5.1.1. Вычисление отклонений от среднего значения



Определение и формула для стандартного отклонения и дисперсии

Стандартное отклонение определяется как величина, которая вычисляется следующим образом. Обратите внимание на то, что при этом вычисляется также **дисперсия** (квадрат стандартного отклонения). Дисперсию иногда используют в качестве меры изменчивости в статистике, особенно когда работают непосредственно с формулами (как вы увидите в главе 15 при рассмотрении *дисперсионного анализа — analysis of variance*, или *ANOVA*). Однако часто в качестве меры изменчивости лучше брать стандартное отклонение. Дисперсия не несет никакой дополнительной (по сравнению со стандартным отклонением) информации, и в то же время, в практических применениях ее сложнее интерпретировать, чем стандартное отклонение. Так, например, в случае набора данных, содержащего потраченные суммы денег (измеренные в долларах), дисперсия будет выражаться в "долларах в квадрате", — это единица измерения, которую трудно себе представить; в то же время стандартное отклонение для этого набора данных будет выражено в привычных для всех долларах.

[4] Путем алгебраических преобразований можно показать, что сумма отклонений от среднего для любого набора данных всегда будет равна нулю. Казалось бы, следует просто заменить знаки "минус" на "плюс" и затем провести усреднение. Однако несложно также показать, что такой простой метод не обеспечивает эффективного использования всей содержащейся в данных информации, если речь идет о нормальном распределении.

[5] Деление на n–1 вместо n (как это обычно делают при вычислении среднего значения) связано с поправкой, обусловленной тем фактом, что при работе с выборкой истинное значение среднего генеральной совокупности неизвестно. Можно также считать, что эта поправка обусловлена потерей при вычислении отклонений одной порции информации (или, как говорят в статистике, одной степени свободы). Потерянной является информация об истинных значениях данных (поскольку теперь, при работе с отклонениями, данные расположены не вокруг среднего, а вокруг нуля).

Вычисление стандартного отклонения для выборки

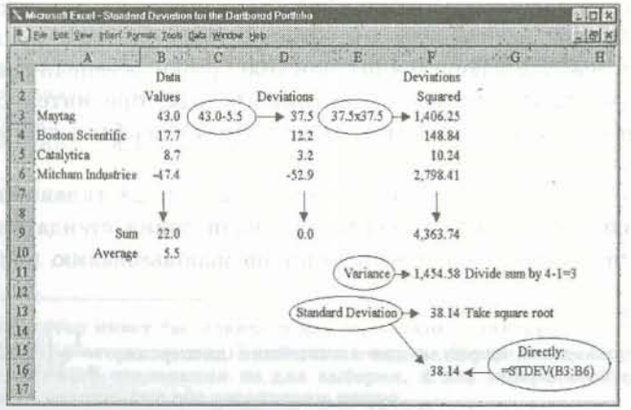
Для того чтобы найти стандартное отклонение для выборки, необходимо выполнить следующие действия.

1. Найти отклонения, вычитая из каждого значения среднее.
2. Возвести полученные значения в квадрат, сложить и разделить полученную сумму на n–1. Результатом будет дисперсия.
3. Извлечь из полученного значения квадратный корень. Это и будет стандартное отклонение.

В табл. 5.1.2 описанная выше процедура проиллюстрирована на примере акций компаний, выбранных выше случайным образом. В результате деления суммы возведенных в квадрат отклонений, 4363,74, на n–1 получаем дисперсию 4363,74 / 3 = 1454,58. Извлекая квадратный корень, получаем стандартное отклонение 38,14. Это значение действительно является разумным описанием собственно отклонений (если не учитывать знаки и рассматривать в первую очередь *величину* отклонений). Последняя формула в правом нижнем углу показывает, как можно вычислить стандартное отклонение непосредственно за один шаг в электронных таблицах Excel®.

Формула для вычисления стандартного отклонения является краткой математической записью описанной выше процедуры. Стандартное отклонение для

Таблица 5.1.2. Вычисление суммы квадратов отклонений, дисперсии и стандартного отклонения



выборки данных обозначается буквой S, и формулы вычисления стандартного отклонения и дисперсии имеют следующий вид[6].

Стандартное отклонение для выборки

$$S = \sqrt{\frac{\text{Сумма квадратов отклонений}}{\text{Количество элементов в выборке} - 1}} =$$

$$= \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}} =$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Дисперсия для выборки

$$\text{Дисперсия} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Вычисление стандартного отклонения для нашего простого примера с использованием соответствующей формулы дает тот же результат, 38.14:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} =$$

$$= \sqrt{\frac{4363,74}{4-1}} =$$

$$= \sqrt{1454,58} =$$

$$= 38,14.$$

Использование калькулятора или компьютера

Конечно, существует другой, значительно более простой способ вычисления стандартного отклонения: использовать калькулятор или компьютер. Именно так *действительно* вычисляют стандартное отклонение, возлагая проведение всех вычислений на специально предназначенное для этого электронное устройство. Однако действия, с которыми мы ознакомились выше, и приведенные выше формулы для нас все равно важны, поскольку они обеспечивают понимание тех чисел, которые вычисляются автоматически. Так, при интерпретации стандартного отклонения важно понимать, что это типичная (или *стандартная*) величина отклонения.

Если на вашем калькуляторе есть клавиша Σ или Σ+ (клавиша суммирования), его, видимо, можно использовать для вычисления стандартного отклонения. За деталями обратитесь к инструкции по использованию калькулятора, в

[6] Другая часто используемая формула для вычисления дисперсии, $\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$, дает тот же результат.

которой, скорее всего, будет описано следующее. Прежде всего необходимо очистить регистры памяти калькулятора, чтобы подготовить его к новым вычислениям. Затем следует ввести в калькулятор последовательность всех значений, нажимая после ввода каждого значения клавишу суммирования. После этого для вычисления стандартного отклонения необходимо нажать клавишу, скорее всего, обозначенную S или *сигма* (σ) [7].

Выбор одного из многих различных способов расчета стандартного отклонения с помощью компьютера зависит от того, какое программное обеспечение используется: электронные таблицы, программы для работы с базами данных, компиляторы языков программирования или специальные программы для статистического анализа данных.

Интерпретация стандартного отклонения

Стандартное отклонение имеет простую и понятную интерпретацию: эта величина описывает *типичное расстояние от среднего значения* для отдельных значений набора данных. Таким образом, стандартное отклонение выступает в качестве меры изменчивости для этих отдельных значений. Поскольку стандартное отклонение отражает типичную величину отклонения, то можно предположить, что для одних значений отклонение будет меньше стандартного, а для других — больше. Таким образом, мы можем ожидать, что для некоторых значений их расстояния от среднего будут меньше стандартного отклонения, в то время как для других значений это расстояние будет превышать величину стандартного отклонения.

На рис. 5.1.1 показано, как можно изобразить стандартное отклонение в виде расстояния от среднего значения. Поскольку среднее показывает центр всего набора данных, отдельные значения будут, по-видимому, располагаться по обе стороны от среднего.



Рис. 5.1.1. Числовая ось с показанными на ней средним значением и стандартным отклонением. Обратите внимание на то, что среднее — это некоторая фиксированная точка на числовой оси (она показывает абсолютную величину), а стандартное отклонение задает определенное расстояние, а именно типичное расстояние от среднего значения

[7] Если ваш калькулятор имеет две клавиши для вычисления стандартного отклонения, одна из которых помечена n, а вторая — n – 1, выберите вторую. Другая клавиша предназначена для вычисления стандартного отклонения не для выборки, а для генеральной совокупности. Различие между этими величинами мы рассмотрим позже.

Пример. *Затраты на рекламу*

Предположим, что ваша фирма тратит 19 миллионов долларов в год на рекламу и руководство фирмы желает знать, соответствует ли эта сумма реальным потребностям. Несмотря на то что существует довольно много способов оценки этой стратегически важной величины, всегда полезно сравнить себя с конкурентами. Пусть другие работающие в вашей отрасли фирмы, имеющие приблизительно такой же размер, в среднем тратят в год на рекламные цели 22,3 миллиона долларов Вы можете воспользоваться стандартным отклонением для того, чтобы исходя из разницы (22,3 – 19 = 3,3 миллиона долларов) оценить, насколько затраты на рекламу вашей фирмы меньше, чем в других аналогичных фирмах.

Рассмотрим затраты на рекламу (в миллионах долларов) группы из n = 17 фирм, похожих на вашу:

8; 19; 22; 20; 27; 37; 38; 23; 12; 11; 23; 20; 18; 23; 35; 11.

Легко убедиться, что среднее составляет 22,3 миллиона долларов (результат округления величины 22,29411 миллиона долларов.) и стандартное отклонение равно 9,18 миллиона долларов (результат округления значения 9,177177).

Поскольку разница между затратами на рекламу в вашей фирме и средними затратами на рекламу в группе фирм (3,3 миллиона долларов) даже меньше одного стандартного отклонения (9,18 миллиона долларов), то можно сделать вывод, что бюджет рекламной деятельности вашей фирмы достаточно типичен. Несмотря на то что он меньше среднего значения, он ближе к этому среднему, чем бюджет типичной фирмы из рассматриваемой группы.

Представим положение вашей фирмы в группе более наглядно. На рис. 5.1.2 вы видите гистограмму размеров затрат на рекламу фирм рассматриваемой группы. На этой гистограмме приведены среднее значение и стандартное отклонение (обратите внимание на то, насколько действительно эффективно стандартное отклонение отражает величину разброса данных по обе стороны от среднего значения). Ваша фирма с бюджетом рекламы в 19 миллионов долларов действительно оказывается достаточно типичной. Несмотря на то что разность в 3,3 миллиона долларов между бюджетом вашей фирмы и средним значением кажется в денежном выражении довольно значительной, она невелика в сравнении с индивидуальными различиями, существующими между бюджетами фирм, входящих в рассматриваемую группу. С точки зрения объема бюджета рекламы положение вашей фирмы не намного ниже среднего.

Пример. *Учет различий между клиентами*

Ваши клиенты отличаются друг от друга; между ними существуют различия в размерах заказов, предпочтении различных товаров, цикличности работы в течение года, потребности в информации, приверженности работе с вами и т.п. Однако у вас, видимо, есть определенное суждение о "типичном клиенте", а также некоторое представление о степени различий между клиентами.

Вы можете обобщить информацию о заявках клиентов с помощью среднего и стандартного отклонения.

Общий годовой объем заказа на одного клиента	
Среднее значение	$85600
Стандартное отклонение	$28300

Таким образом, за последний год каждый клиент в среднем заказал товаров на сумму $85600. В качестве характеристики различий между клиентами выступает стандартное отклонение. Его величина в $28300 показывает, что обычно ваши клиенты делали заказы на суммы, меньшие или большие примерно на $28300, чем среднее значение в $85600. Слово "примерно" имеет здесь большое значение: некоторые клиенты могут делать заказы, размер которых очень близок к среднему уровню, в то время как для других будут наблюдаться отличия, значительно превышающие $28300. Среднее значение показывает типичный объем поступивших в течение года заказов от одного клиента, а стандартное отклонение иллюстрирует типичное отклонение от среднего.

Обратите внимание также на то, что стандартное отклонение измеряется в тех же единицах, что и среднее значение; в данном случае обе эти величины определены в долларах. Если говорить точнее, то единица измерения здесь — это "доллар в год на одного клиента". Это увязывает единицы измерения с исходным набором данных, который представляет собой последовательность объемов "долларов в год", причем каждому клиенту соответствует одно число.

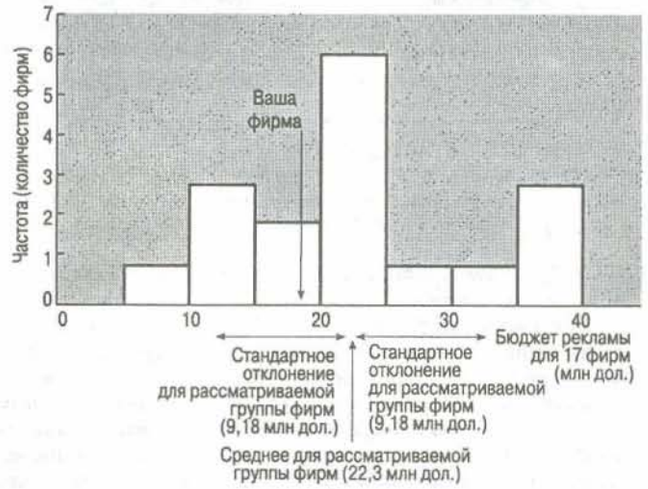


Рис. 5.1.2. Гистограмма размеров затрат на рекламу для группы 17 подобных фирм. Показаны среднее значение и стандартное отклонение. Бюджет вашей фирмы, составляющий 19 миллионов долларов, достаточно типичен на фоне аналогичных фирм. Он отличается от среднего даже меньше, чем на одну величину стандартного отклонения

Интерпретация стандартного отклонения для нормального распределения

В том случае, когда набор данных имеет приблизительно нормальное распределение, стандартное отклонение приобретает особый смысл. Приблизительно две трети значений из такого набора данных находятся *в пределах одного стандартного отклонения* по обе стороны от среднего значения, как показано на рис. 5.1.3.

Так, например, если способности ваших работников распределены приблизительно нормально, то вы можете ожидать, что оценки способностей примерно двух третей из них попадают на расстояния не более одной величины стандартного отклонения от среднего значения — либо выше, либо ниже среднего. Это означает, что приблизительно треть работников имеет способности, лежащие в пределах одной величины стандартного отклонения выше среднего, а примерно треть — в соответствующей области ниже среднего. Остальные работники, которых также приблизительно треть, распределятся таким образом: около половины этой одной трети (шестая часть всех работников) обладает способностями, превышающими средние более чем на величину одного стандартного отклоне-

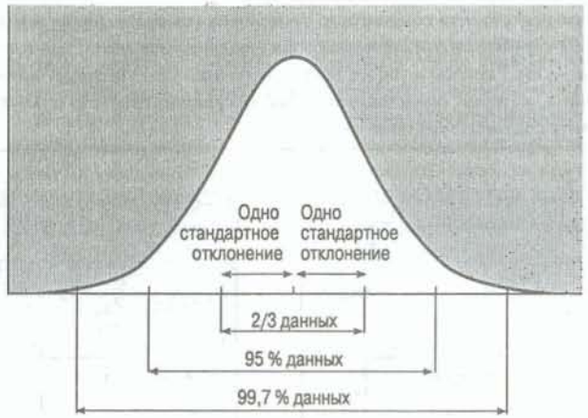


Рис. 5.1.3. В случае нормального распределения значения в наборе данных можно легко разделить в зависимости от выраженного в значениях стандартного отклонения их расстояния от среднего значения. Примерно две трети всех значений находятся не далее одного стандартного отклонения от среднего. Около 95% всех значений находятся в пределах двух стандартных отклонений от среднего. И, наконец, мы можем ожидать, что обнаружим почти все данные (99,7%) на расстоянии не более трех стандартных отклонений от среднего

ния, и примерно шестая часть всех работников (увы!) окажется ниже среднего далее, чем на расстоянии одного стандартного отклонения.

Из рис. 5.1.3 также видно, что в случае нормального распределения следует ожидать, что примерно 95% всех данных окажутся *в пределах двух величин стандартного отклонения от среднего значения*[8]. Этот факт будет иметь большое значение при рассмотрении статистических выводов, поскольку допустимые погрешности оценок часто ограничиваются величиной 5%.

И, наконец, мы вправе предположить, что почти все данные (99,7%) будут находиться *в пределах трех величин стандартного отклонения от среднего значения*. При этом только 0,3% всех значений набора данных оказываются от среднего на большем удалении. На рис. 5.1.3 можно видеть, что график нормального распределения на расстояниях порядка трех стандартных отклонений от среднего опускается почти до нуля. В картах контроля, которые широко используют для контроля качества продукции, пределы часто устанавливаются таким образом, чтобы в качестве заслуживающей внимания проблемы выступали именно те результаты наблюдений, которые отстоят от среднего на расстоянии, большем чем три стандартных отклонения.

Что же происходит в том случае, если набор данных не подчиняется нормальному распределению? В таком случае описанные выше проценты применять нель-

[8] В случае идеального нормального распределения в точности 95% всех данных попадают в область вблизи среднего значения в пределах 1,96 стандартного отклонения. Поскольку величина 1,96 достаточно близка к значению 2, мы используем описание "две величины стандартного отклонения" в качестве удобного и хорошего приближения.

зя. К сожалению, поскольку существует множество скошенных (или других, отличающихся от нормального) распределений, нельзя указать единое правило определения таких процентов для произвольного распределения[9]. На рис. 5.1.4 приведен пример скошенного (асимметричного) распределения. В этом случае на расстоянии, не превышающем одно стандартное отклонение, находится не две трети, а три четверти всех данных. Кроме того, большинство этих данных расположено слева от среднего (поскольку здесь график распределения проходит выше).

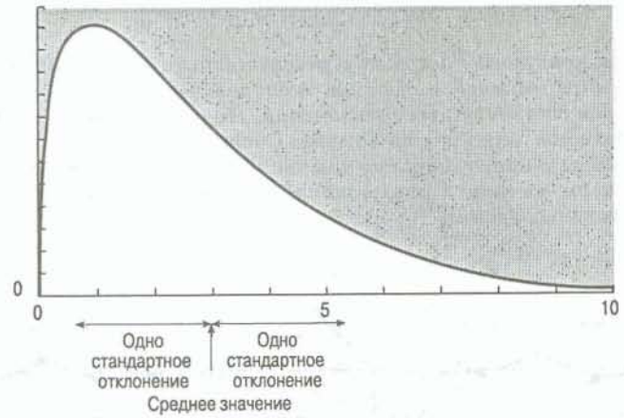


Рис. 5.1.4. В случае скошенного распределения не существует простых правил для определения части данных, попадающих в пределы одного (двух или трех) стандартных отклонений от среднего значения

Пример. *Контрольные карты для контроля качества изображений*

Предприятие выпускает экраны мониторов и использует для контроля и улучшения качества продукции карты контроля качества (или, как еще говорят, контрольные карты). В частности, размер одной точки экрана должен быть настолько мал, чтобы пользователь мог видеть мельчайшие детали изображения. Карта контроля содержит результаты отдельных измерений размера точки (которые отличаются для разных мониторов), средние значения (которые, как можно видеть, проходят через центр данных) и контрольные границы (которые устанавливаются выше и ниже среднего значения на расстоянии трех стандартных отклонений; более детально карты контроля качества будут описаны в главе 18). На рис. 5.1.5 показан пример карты контроля качества для набора устройств, у которых результаты всех измерений находятся в пределах контрольных границ. На рис. 5.1.6 приведен другой пример, который демонстрирует, что у монитора №22 наблюдается выход за контрольные границы. Карты контроля качества помогают выявить проблему. Дальнейшее исследование и исправление ситуации зависит от менеджера.

[9] Существует, однако, ограничение, которое называется правилом Чебышева. В соответствии с этим правилом, по меньшей мере $1 - 1/a^2$ значений попадает в промежуток, лежащий в пределах a стандартных отклонений от среднего значения. Например, при a = 2 по меньшей мере 75% данных (это значение рассчитывается как $1 - 1/a^2$) должно находиться не далее, чем на расстоянии удвоенного стандартного отклонения от среднего, даже если распределение не является нормальным (сравните с величиной для нормального распределения, составляющей примерно 95%). Если a = 3, по меньшей мере 88,9% данных будет находиться в пределах утроенного стандартного отклонения от среднего значения.

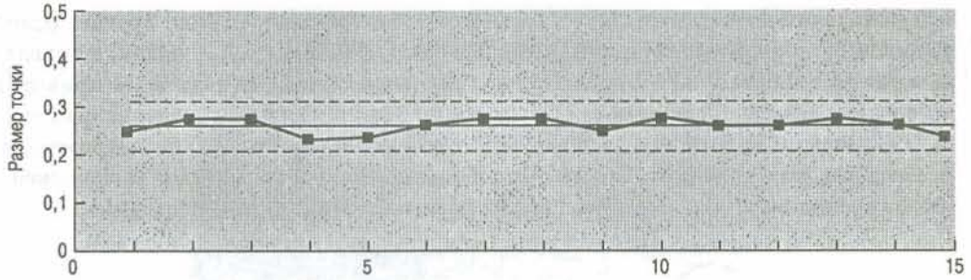


Рис. 5.1.5. Карта контроля качества с результатами измерений для экранов мониторов. Линии нижней и верхней контрольных границ проведены на расстоянии трех стандартных отклонений. Показано также среднее значение, которое проходит через центр данных. Система находится под контролем, и есть только случайные отклонения от среднего, поскольку в отклонениях нет четких тенденций и результатов измерений, которые выходят за пределы контрольных границ

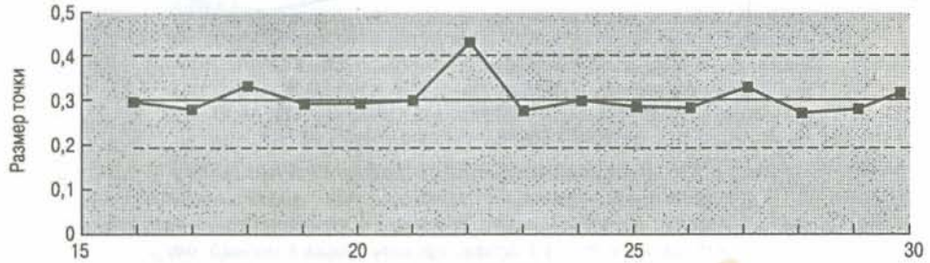


Рис. 5.1.6. Эта система вышла из-под контроля. Обратите внимание на то, что результат измерения для экрана №22 отклоняется от среднего значения более чем на три стандартных отклонения. Это выходит за пределы обычной вариации для системы, возникает необходимость в изучении причин и устранении подобных проблем в дальнейшем

Пример. *Размеры прибыли на бирже меняются каждый день*

В этом примере мы исследуем непостоянство фондовой биржи (измеренное соответствующим стандартным отклонением) за период времени, предшествовавший обвалу в 1987 году. В табл. 5.1.3 приведены цены на фондовой бирже, зафиксированные в момент закрытия каждого из рабочих дней с 31 июля до 9 октября 1987 года. Цены измерены с помощью индекса Доу Джонс (*Dow Jones Industrial*). Индекс Доу Джонс вычисляется как среднее значение рыночных цен акций 30 специально отобранных крупных промышленных компаний. Обычно инвесторы изучают такие данные в виде графика зависимости индекса цен от времени. Такой график показан на рис. 5.1.7.

Вместо биржевых цен финансовые аналитики и исследователи часто рассматривают *дневную прибыль*, представляющую собой процентную ставку, которую можно получить при инвестировании в акции за одни сутки. Эта величина вычисляется путем деления изменения индекса за сутки на его величину в предыдущий день. Например, дневная прибыль за 3 августа

$$\frac{(2557{,}08-2572{,}07)}{2572{,}07}=-0{,}006$$

представляет собой потери размером несколько менее 1%.[10] Показатель дневной прибыли отражает динамику происходящих ежедневно изменений на рынке в более явном виде, чем собственно величина среднего значения цен. В табл. 5.1.4 приведена динамика изменения дневной прибыли.

Таблица 5.1.3 Цены фондовой биржи

Индекс Доу Джонс	Дата	Индекс Доу Джонс	Дата
2572,07	31 июля 1987 г.	2561,38	
2557,08		2545,12	
2546,72		2549,27	
2566,65		2576,05	
2594,23		2608,74	
2592,00		2613,04	
2635,84		2566,58	
2680,48		2530,19	
2669,32		2527,90	
2691,49		2524,64	
2685,43		2492,82	
2700,57		2568,05	
2654,66		2585,67	
2665,82		2566,42	
2706,79		2570,17	
2709,50		2601,50	
2697,07		2590,57	
2722,42		2596,28	
2701,85		2693,20	
2675,06		2640,99	
2639,35		2640,18	
2662,95		2548,63	
2610,97		2551,08	
2602,04		2516,64	
2599,49		2482,21	9 октября 1987 г.

Источник: *Daily Stock Price Record, New York Stock Exchange*, Standard & Poor's Corporation, 1987.

[10] Здесь указан доход за период с 31 июля по 3 августа. Мы считаем его дневным доходом, поскольку два входящих в этот промежуток времени дня были выходными и в эти дни торговля на бирже не велась.

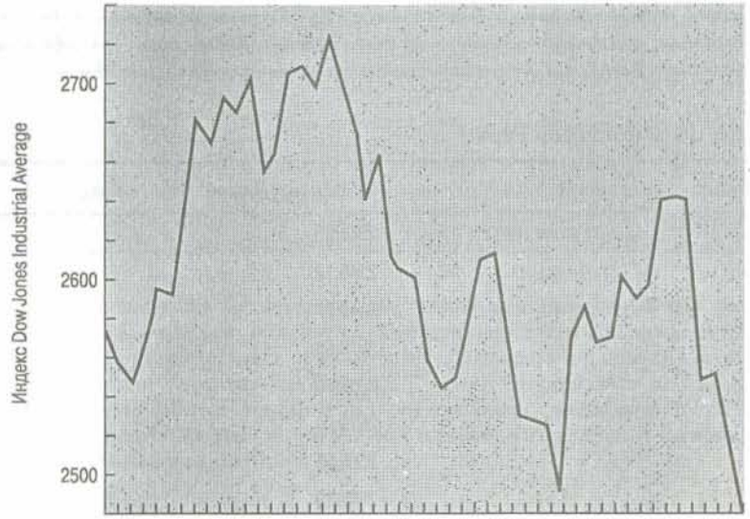


Рис. 5.1.7. Значение индекса Доу Джонса на момент закрытия торгов на бирже в период с 31 июля по 9 октября 1987 года

Таблица 5.1.4. Дневная прибыль на бирже

Индекс Доу Джонса	Дневная прибыль	Индекс Доу Джонса	Дневная прибыль
2572,07 (7/31/87)		2561,38	–0,015
2557,08 (8/3/87)	–0,006 = (2557,08 – 2572,07) / 2572,07	2545,12	–0,006
2546,72 (8/4/87)	–0,004	2549,27	0,002
2566,65 (8/5/87)	0,008	2576,05	0,011
2594,23	0,011	2608,74	0,013
2592,00	–0,001	2613,04	0,002
2635,84	0,017	2566,58	–0,018
2680,48	0,017	2530,19	–0,014
2669,32	–0,004	2527,90	–0,001
2691,49	0,008	2524,64	–0,001
2685,43	–0,002	2492,82	–0,013
2700,57	0,006	2568,05	0,030
2654,66	–0,017	2585,67	0,007
2665,82	0,004	2566,42	–0,007
2706,79	0,015	2570,17	0,001
2709,50	0,001	2601,50	0,012
2697,07	–0,005	2590,57	–0,004
2722,42	0,009	2596,28	0,002

Индекс Доу Джонса	Дневная прибыль	Индекс Доу Джонса	Дневная прибыль
2701,85	–0,008	2693,20	0,017
2675,06	–0,010	2640,99	0,001
2639,35	–0,013	2640,18	–0,000
2662,95	0,009	2548,63	–0,035
2610,97	–0,020	2551,08	0,001
2602,04	–0,003	2516,64	–0,014
2599,49	–0,001	2482,21 (10/9/87)	–0,014

На рис. 5.1.8 показана гистограмма дневной прибыли. Она свидетельствует о том, что приведенные значения имеют нормальное распределение. Средняя дневная прибыль за этот период времени составляла –0,0007, т.е. она была приблизительно равна нулю (среднее снижение составило семь сотых процента). Таким образом, на рынке в это время держался средний курс. Стандартное отклонение составляет 0,0117. Это означает, что $1, вложенный в фондовый рынок, в среднем изменялся за сутки на $0,0117, в том смысле, что вложение $1 могло привести за сутки к прибыли или потере примерно в $0,0117.

Крайние значения, которые мы видим на рис. 5.1.8 с обеих сторон от центра, демонстрируют максимальный размер роста и падения за один день. Так, 22 сентября на рынке наблюдался подъем с 2492,82 до 2568,05, что составило рост на 75,23 пункта, с дневной прибылью 0,030 (прибыль в размере $0,030 на один доллар, вложенный днем раньше). А 6 октября на рынке произошло понижение с 2640,18 до 2548,63, т.е. на 91,55 пункта. Дневная прибыль при этом составила –0,035 (потери в размере 0,035 доллара на один доллар, вложенный днем раньше).

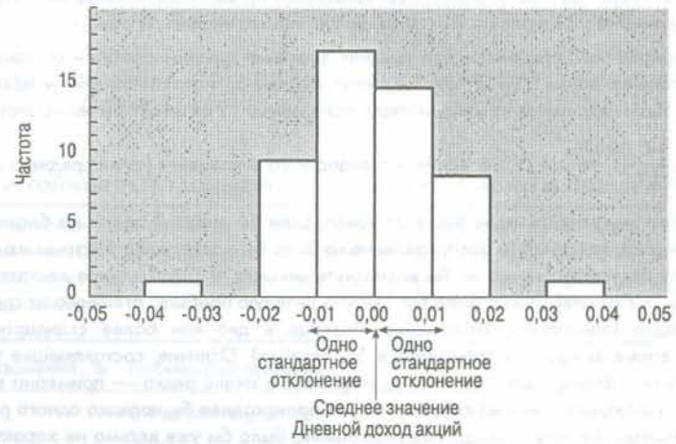


Рис. 5.1.8. Гистограмма дневной прибыли акций на бирже. Среднее значение прибыли близко к нулю. Это означает, что кратковременные повышения и понижения были примерно равновероятными. Стандартное отклонение, составляющее 0,0117, отражает величину обычных суточных флуктуаций. В течение этого времени вложенный на фондовом рынке доллар мог измениться за день на величину порядка одного цента

Для того чтобы оставаться в пределах одной величины стандартного отклонения (0,0117) от среднего значения (–0,0007), размер дневной прибыли должен находиться в пределах от –0,0007 – 0,0117 = –0,0124 до –0,0007 + 0,0117 = 0,0110. Из 49 приведенных значений дневной прибыли этому требованию отвечают 32. Таким образом, 32/49, или 65,3% значений дневной прибыли удалены от среднего значения на расстояние, не превышающее одного стандартного отклонения. Этот процент достаточно близок к значению 2/3 (или 66,7%) — приблизительной той части от общего количества значений, которую мы могли бы ожидать в случае идеального нормального распределения. Следовательно, *можно считать*, что "правило двух третей" работает.

Для того чтобы оставаться в пределах расстояния в две величины стандартного отклонения от среднего значения, дневная прибыль должна находиться в пределах от –0,0007 – (2×0,0117) = –0,0241 до –0,0007 + (2×0,0117) = 0,0227. Из 49 значений дневной прибыли этому требованию отвечают 47 (все, за исключением двух крайних значений, на которые мы уже обратили внимание ранее). Таким образом, 47/49, или 95,9%, величин дневной прибыли расположены по отношению к среднему значению на расстоянии, не превышающем удвоенного стандартного отклонения. *Полученное значение достаточно близко к значению 95%, которое мы могли бы ожидать в случае идеального нормального распределения.*

Этот пример достаточно хорошо соответствует правилам для нормального распределения. В других примерах распределений, также близких к нормальному, не удивительным будет наличие больших расхождений со значениями 2/3, или 95%, которые мы ожидаем получить для идеального нормального распределения.

Пример. *Обвал на фондовой бирже в 1987 году: 19 стандартных отклонений!*

В понедельник 19 октября 1987 года индекс цен Доу Джонса упал на 508 пунктов, с 2246,74 (в предыдущую пятницу) до 1738,74. Это соответствует дневному доходу –0,2261; таким образом, фондовый рынок потерял 22,61% своей стоимости. Это неожиданное падение стоимости, показанное на рис. 5.1.9, было самым большим со времени "Большого кризиса" 1929 года.

Для того чтобы представить себе, насколько экстремальной с точки зрения статистики оказалась ситуация при этом обвале, сравним ее с тем, что следовало бы ожидать в соответствии с предшествовавшим поведением рынка. В качестве базового периода воспользуемся предыдущим примером, в котором рассмотрен промежуток времени с 31 июля по 9 октября, до пятницы за неделю до обвала.

Для базового периода мы определили, что среднее значение дневной прибыли составляет –0,0007, а стандартное отклонение равно 0,0117. Сколько величин стандартного отклонения необходимо отложить вниз от среднего значения, чтобы получить потери, понесенные 19 октября? Ответ на этот вопрос таков:

$$\frac{-0,2261-(-0,0007)}{0,0117}=-19,26$$ величин стандартного отклонения (ниже среднего значения).

Отсюда видно, насколько необычным был этот обвал. Если бы дневной доход на бирже действительно имел нормальное распределение (и распределение не было бы подвержено быстрым изменениям), такого экстремального результата не могло бы возникнуть *никогда*. В таком случае достаточно часто (примерно в одной трети случаев) можно было бы ожидать дневную прибыль, отличную от среднего значения более чем на одно стандартное отклонение. Разница в два или более стандартных отклонений наблюдалось бы время от времени (примерно в 5% случаев). Отличие, составляющее утроенное стандартное отклонение и более, могло бы наблюдаться только очень редко — *примерно в 0,3% случаев*, или, для большей наглядности, можно сказать, что это происходило бы порядка одного раза в год[11]. Даже отличие, составляющее пять стандартных отклонений, было бы уже весьма не характерным для идеального нормального распределения. Разница в 19,26 стандартных отклонений представляется в таком случае просто практически невероятной.

Однако, несмотря на это, мы уже увидели дневную прибыль, отличную от среднего значения на 19,26 величин стандартного отклонения. Это показывает, что размер дневной прибыли на фондовой бирже *не* подчиняется идеальному нормальному распределению. Это не означает, что теория в чем-то неверна; это указывает лишь на то, что теория в данном случае неприменима. Несмотря на то что нормальное распределение описывает дневную прибыль для большей части времени работы фондовой биржи, обвал 1987 года напоминает нам о необходимости проверки правильности всех используемых предположений для защиты своих интересов в особых случаях.

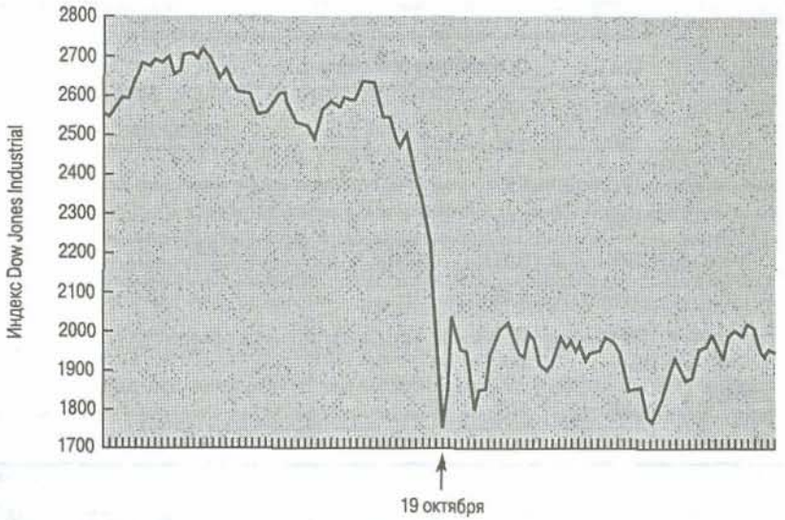


Рис. 5.1.9. Значения индекса Доу Джонса в момент закрытия торгов на бирже в период с 31 июля до 31 декабря 1987 года

Пример. *Неустойчивость фондового рынка до обвала и после него*

В период, последовавший за обвалом 19 октября 1987 года, состояние рынка характеризовалось высокой неустойчивостью. Степень неустойчивости можно оценить, воспользовавшись, как и в предыдущем примере, стандартным отклонением дневной прибыли. Ниже приведена таблица значений стандартного отклонения.

Стандартное отклонение, %	Промежуток времени
1,17	С 1 августа по 9 октября
8,36	С 12 октября (за 1 неделю до обвала) по 26 октября (1 неделя после обвала)
2,09	С 27 октября по 31 декабря 1987 года

В период времени, непосредственно примыкающий к обвалу (одна неделя до него и одна неделя после), стандартное отклонение было приблизительно в семь раз выше, чем до этого периода. После обвала стандартное отклонение уменьшилось, однако осталось примерно в два раза выше, чем до него (2,09% по сравнению с 1,17%). Рынок после обвала, безусловно, в значительной мере "вернулся к деловой ак-

[11] То, что событие, которое наблюдается только в 0,3% всех дней, будет происходить примерно раз в год, можно определить из следующих рассуждений. Во-первых, 0,3% равно числу 0,003. Во-вторых, обратная величина составляет 1/0,003 = 333 (приблизительно), что означает, что такое событие будет происходить примерно один раз в каждые 333 дня или (очень приблизительно) один раз в год.

тивности", однако он остался "неспокойным", о чем свидетельствует высокая неустойчивость, которая измеряется стандартным отклонением.

Повышение неустойчивости отображено на рис. 5.1.10. Если абстрагироваться от сильных колебаний рынка накануне и сразу после 19 октября, можно увидеть, что размах по вертикали колебаний графика справа от этой даты примерно в два раза выше, чем слева от нее. Эти оценки неустойчивости отражены и в приведенной выше таблице стандартных отклонений, примерно соответствующих показанным на графике расстояниям по вертикали.

В последнее время неустойчивость фондового рынка значительно снизилась. Ниже приведена таблица стандартных отклонений дневной прибыли для каждого года с 1990 по 1998, рассчитанная для фондового индекса *S&P 500*. Обратите внимание, что типичное изменение цен в 1995 году составляло примерно половину процента (от стоимости всего портфеля) в день, однако затем неустойчивость рынка стала расти.

Год	Стандартное отклонение, %
1990	1,00
1991	0,89
1992	0,60
1993	0,53
1994	0,61
1995	0,48
1996	0,73
1997	1,12
1998	1,26

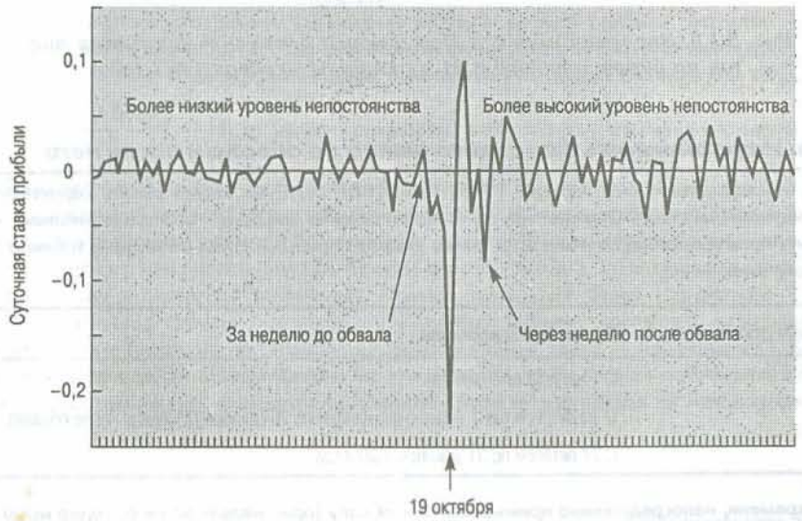


Рис. 5.1.10. Дневная прибыль в период с 1 августа по 31 декабря 1987 года. Обратите внимание на повышение неустойчивости рынка после обвала

Пример. *Диверсификация на фондовом рынке*

Покупая акции, вы всегда рискуете, поскольку с течением времени курс акций может либо вырасти, либо понизиться. Владение акциями сразу нескольких различных компаний называется *диверсификацией*. Диверсификация снижает риск, поскольку владелец портфеля акций нескольких компаний в меньшей степени подвержен влиянию возможных резких изменений курса акций конкретной компании. Рассмотрим величины риска для трех случаев: (1) владение только акциями *Boeing*, (2) владение только акциями *Johnson & Johnson* и (3) владение портфелем из названных акций в равных долях. Стандартное отклонение дневной ставки прибыли для каждого из этих случаев (за 1994 год и три первых квартала 1995 года) приведено в следующей таблице.

Портфель	Стандартное отклонение, %
Johnson & Johnson	1,39
Boeing	1,46
Обе компании,	0,99

Обратите внимание на снижение риска в случае владения акциями более чем одной компании (риск снижается примерно с 1,4% в день до величины порядка 1% в день). Если портфель содержит акции большего количества компаний, риск удается снизить еще больше. Рискованность акций *S&P 500* (включающих акции 500 различных компаний) была в этот период еще меньше — порядка 0,6%.

Стандартное отклонение выборки и генеральной совокупности

Существуют два различных (однако связанных между собой) вида стандартного отклонения: **стандартное отклонение выборки** (для выборки, сделанной из большей генеральной совокупности, обозначается буквой S) и **стандартное отклонение генеральной совокупности** (для всей генеральной совокупности, обозначается буквой σ — малая греческая буква "сигма").

Названия этих величин отражают правила их использования. В случае работы с выборкой данных, взятых случайным образом из большей генеральной совокупности, используется стандартное отклонение выборки. Если же изучается вся генеральная совокупность, следует использовать стандартное отклонение генеральной совокупности (часто также используют термины "выборочное стандартное отклонение" и "генеральное стандартное отклонение" соответственно. — *Прим. ред.*). Стандартное отклонение выборки несколько больше, что обеспечивает поправку на случайность самой выборки.

В некоторых случаях ситуация может быть не совсем однозначной. Так, например, набор данных о заработной плате всех сотрудников некоторой компании можно рассматривать и как генеральную совокупность (поскольку рассматриваются все работники этой компании) и как выборку (если рассматривать сотрудников компании как представителей большей генеральной совокупности подобного рода специалистов). Такая неоднозначность является следствием оценки рассматриваемой ситуации, а не следствием характера самих данных. Если считать, что данные охватывают полностью круг решаемых задач, то эти данные, безусловно, представляют собой генеральную совокупность. Если же мы ставим цель провести некоторое обобщение (например, перейти от рассмотрения сотрудников данной компании к рассмотрению сотрудников, работающих в ана-

логичных компаниях), то те же данные можно считать выборкой из некоторой (возможно, гипотетической) генеральной совокупности.

Чтобы покончить с оставшимися неясностями, примем следующее правило: *при наличии сомнений использовать стандартное отклонение для выборки*. Эта величина несколько больше, и выбрать ее — значит, поступить более осторожно и консервативно, и в конечном итоге не допустить систематической недооценки неопределенности.

Что касается вычислений, то единственное различие между этими двумя показателями состоит в том, что при вычислении стандартного отклонения выборки вычитают 1 (т.е. делят на n−1), а при вычислении стандартного отклонения генеральной совокупности не вычитают 1 (т.е. делят на N). В связи с этим использование формулы стандартного отклонения для выборки дает несколько большее значение для небольших размеров выборки, что отражает увеличение неопределенности, обусловленной использованием выборки вместо всего множества данных[12]. Существуют также некоторые общепринятые различия в обозначениях. Среднее для выборки из n элементов обозначается $\bar{X}$, а среднее генеральной совокупности из N элементов обозначается греческой буквой μ ("мю"). Формулы для расчета стандартного отклонения имеют следующий вид.

Стандартное отклонение для выборки

$$s = \sqrt{\frac{\text{Сумма квадратов отклонений}}{\text{Количество элементов в выборке} - 1}} =$$

$$= \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}} =$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Стандартное отклонение для генеральной совокупности

$$\sigma = \sqrt{\frac{\text{Сумма квадратов отклонений}}{\text{Количество элементов генеральной совокупности}}} =$$

$$= \sqrt{\frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_n - \mu)^2}{N}} =$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}.$$

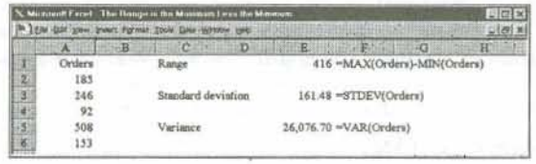
Чем меньше количество элементов (n или N), тем сильнее проявляются различия между этими двумя формулами. В случае 10 элементов стандартное отклонение выборки превышает стандартное отклонение генеральной совокупности на 5,4%. При 25 элементах различие составляет 2,1%. С увеличением числа элементов это расхождение уменьшается, доходя до 1,0% для 50 элементов и 0,5% для 100 элементов. Таким образом, в случае достаточно большого количества данных различие между двумя рассмотренными методами вычислений оказывается несущественным.

[12] Справедливо также следующее утверждение: при делении не на n, а на n − 1 дисперсия выборки (квадрат стандартного отклонения) становится "несмещенной" оценкой (т.е. корректной и для генеральной совокупности в среднем). Однако стандартное отклонение выборки попрежнему остается "смещенной" оценкой стандартного отклонения генеральной совокупности. Детали построения выборки из генеральной совокупности будут рассмотрены в главе 8.

5.2. Размах: быстрая и поверхностная оценка

Размах, или интервал, занимаемый значениями данных, равен разности между самым большим и самым малым значениями. Он определяет, до какой степени отдельные значения отличаются между собой. Ниже показано вычисление размаха для небольшого набора данных, представляющих количество полученных за последнее время заказов на пять различных видов товара[13]:

Размах набора данных (185, 246, 92, 508, 153) =
= максимальное — минимальное = 508 − 92 = 416.



	A	B	C	D	E	F	G	H
1	Orders		Range		416	=MAX(Orders)-MIN(Orders)		
2	185							
3	246		Standard deviation		161.48	=STDEV(Orders)		
4	92							
5	508		Variance		26,076.70	=VAR(Orders)		
6	153							

Обратите внимание, что размах очень легко вычислить. Для этого нужно только просмотреть список значений, выбрать из списка самое большое и самое малое значения, а затем вычесть из большего меньшее. Раньше, до появления электронных калькуляторов и компьютеров, простота вычисления размаха была причиной того, что этот показатель часто использовался в качестве меры изменчивости. Теперь, когда вычислять стандартное отклонение стало намного проще, размах используют не так часто.

Когда важны экстремальные значения (т.е. наибольшее и наименьшее), размах может быть хорошей мерой разброса. Примером может быть необходимость описать пределы изменения значений данных. Такая характеристика может оказаться полезной для двух целей: во-первых, для *описания* границ изменения данных и, во-вторых, для *поиска ошибок* в значениях. При наличии в наборе данных очень больших (или очень малых) ошибочно записанных значений размах имеет тенденцию возрастать и сразу же становится, с позиций здравого смысла, слишком большим. Такая особенность делает размах полезным для поиска ошибок и *редактирования* значений данных.

С другой стороны, в силу чувствительности к предельным значениям размах оказывается не слишком полезным в качестве такой статистической меры разброса, которая характеризует набор данных в целом. Размах не отражает типич-

[13] Для того чтобы вычисления в Excel® выполнялись по формулам так, как это показано на рисунке, сначала необходимо дать пяти данным числам имя "Заказы" ("Orders"). Для этого необходимо выделить эти пять чисел и выполнить пункт меню Insert⇨Name⇨Define (Вставка⇨Имя⇨Присвоить)

ную изменчивость данных, а скорее фокусируется всего лишь на двух значениях. Стандартное отклонение более чувствительно ко всем данным, благодаря чему эта величина дает более четкое представление об общей картине. Размах *всегда* больше, чем стандартное отклонение.

Пример. *Заработная плата персонала*

Рассмотрим заработную плату наемных работников технического отдела фирмы, производящей товары потребительской электроники. Необходимые нам значения приведены в табл. 5.2.1. Не будем обращать внимание на идентификационные коды и сосредоточимся на столбце заработной платы. Легко увидеть, что самый высокооплачиваемый работник получает в год $44500 (должность руководителя технического отдела), а самый низкооплачиваемый — 16500 дол. (молодой перспективный сотрудник, который еще не закончил образование). Размах составляет $28000 (= 44500 – 16500). Эта величина показывает различие в долларовом выражении между наиболее и наименее оплачиваемыми сотрудниками, как это показано на рис. 5.2.1.

Обратите внимание, что размах рассчитывается исходя из двух предельных значений: максимальной и минимальной зарплаты. Величина размаха не отражает типичную вариацию зарплаты в отделе, для этого необходимо использовать стандартное отклонение.

Для более полного анализа (чтобы удовлетворить интерес читателя или дать ему возможность проверить вычисленные самостоятельно значения) можно указать, что средняя зарплата в отделе составляет $27950 и стандартное отклонение (которое лучше характеризует *типичную* изменчивость зарплаты) равно $8581. Это стандартное отклонение выборки. При этом инженеры фирмы рассматриваются как

Таблица 5.2.1. Зарплата персонала

Идентификационный код сотрудника	Зарплата, дол.	Идентификационный код сотрудника	Зарплата, дол.
918886653	34000	743594601	33000
771631111	26000	731866668	16500
148609912	27000	490731488	44500
742149808	22500	733401899	35000
968454888	21000	589246387	20000

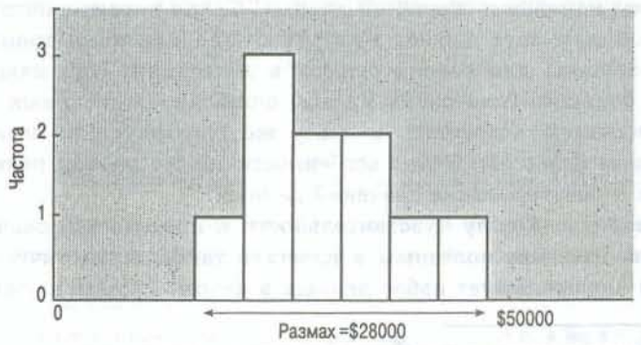


Рис. 5.2.1. Размах заработной платы составляет $28000 (от $16500 до $44500). Эта величина характеризует ширину всей гистограммы

выборка из большого множества всех инженеров, выполняющих аналогичную работу.

Кратко описанную выше ситуацию можно охарактеризовать следующим образом. Величина $28000 (размах) разделяет минимальную и максимальную зарплаты. Стандартное отклонение, $8581, показывает, насколько (приблизительно) зарплаты отдельных сотрудников отличаются от $27950 — средней зарплаты для данной группы.

Пример. *Длительность пребывания пациентов в больнице*

Работа больниц сейчас больше похожа на коммерческую деятельность, чем это было раньше. Отчасти это связано с проникновением конкурентной борьбы в область здравоохранения. Многие организации, оказывающие услуги в области здравоохранения, просто нанимают врачей как служащих, в то время как в традиционных больницах врачи имеют большую независимость. Еще одна причина коммерциализации здравоохранения состоит в том, что согласно программе охраны здоровья *Medicare* в настоящее время есть тенденция производить фиксированные выплаты на основе диагноза, а не гибкие выплаты в зависимости от длительности лечения. Это способствует возникновению сильной тенденции к сокращению длительности лечения в случае конкретной болезни пациента.

В качестве одного из показателей интенсивности лечения выступает количество дней пребывания пациента в больнице. В табл. 5.2.2 отражено количество дней пребывания в больнице для выборки пациентов в прошлом году[14]. Размах этого ряда данных составляет 386–1 = 385 дней, что представляет собой слишком большое значение, поскольку в году только 365 (или 366) дней, а этот набор данных, как предполагается, относится только к одному году. Данный пример иллюстрирует пользу применения понятия размаха для редактирования набора данных с целью выявления ошибок перед началом анализа данных. Для этого также полезно внимательно исследовать наименьшие и наибольшие значения.

При тщательной проверке данных была обнаружена опечатка. Реальное значение 286 было ошибочно записано как 386. В исправленном наборе данных размах стал равен 285 (т.е. 286 – 1).

Таблица 5.2.2. Длительность пребывания в больнице для выборки пациентов

Койко-дней за прошлый год		
17	33	5
16	5	6
1	1	12
1	7	16
7	4	386
74	13	2
2	6	7
163	33	28
51		

[14] Этот гипотетический набор данных основан на опыте одного из моих друзей из центра экономических исследований и проблемах, которые у него возникли, когда он в течение двух недель пытался применить компьютер для анализа записей медицинской статистики при изучении эффективности различных систем предоставления услуг здравоохранения. Успешному исследователю, как в академической области, так и в области коммерции, на пути к более глубокому пониманию часто необходимо преодолевать множество мелких проблем.

5.3. Коэффициент вариации: мера относительной изменчивости

Коэффициент вариации, который определяется как результат деления стандартного отклонения на среднее значение, представляет собой относительную меру изменчивости и выражается в процентах или долях среднего значения. Такой подход особенно полезен в том случае, когда набор данных не содержит отрицательных значений. Формула для вычисления коэффициента вариации имеет следующий вид:

Коэффициент вариации

$$\text{Коэффициент вариации} = \frac{\text{Стандартное отклонение}}{\text{Среднее}}.$$

Для выборки:

$$\text{Выборочный коэффициент вариации} = \frac{S}{\bar{X}}.$$

Для генеральной совокупности:

$$\text{Коэффициент вариации генеральной совокупности} = \frac{\sigma}{\mu}.$$

Обратите внимание на то, что стандартное отклонение стоит в числителе. Таким образом, результат деления характеризует изменчивость.

Так, например, если в среднем покупатель тратит в супермаркете $35,26, а стандартное отклонение составляет $14,08, то коэффициент вариации равен 14,08 / 35,26 = 0,399, или 39,9%. Это означает, что обычно суммы, которые покупатель тратит при посещении супермаркета, отличаются от среднего значения примерно на 39,9%. В абсолютном выражении это типичное отличие от среднего размера затрат равно $14,08 (стандартное отклонение), что составляет 39,9% (коэффициент вариации) от среднего.

Коэффициент вариации — *безразмерная величина*. Это просто число, доля или процент. При вычислении коэффициента вариации размерность исчезает в результате деления стандартного отклонения на среднее значение. Коэффициент вариации полезен в тех случаях, когда важна не абсолютная величина отличий значений данных, но их относительная изменчивость.

Используя коэффициент вариации, можно сравнить вариацию объемов продаж для крупной и малой фирмы "с поправкой на размер фирмы". Обычно у фирмы, оборот которой составляет сотни миллионов долларов, различия в объемах продаж также довольно велики — например, они могут достигать десятков миллионов долларов. Для другой фирмы, объем продаж которой исчисляется миллионами долларов, различия могут составлять сотни тысяч. Однако в каждом из этих двух случаев вариация составляет порядка 10% среднего значения общего объема продаж. Для большей фирмы абсолютное значение вариации окажется больше (большее стандартное отклонение), однако относительная, или учитывающая объем, величина вариации (коэффициент вариации) оказывается одинаковой для обеих фирм.

Следует также отметить, что коэффициент вариации может превысить 100% даже в том случае, если все значения положительны. Это, в частности, может быть в случае сильно скошенного распределения или при наличии значений, сильно отличающихся от среднего. Такой результат означает, что в изучаемой ситуации наблюдается *очень* сильная вариация по отношению к величине среднего значения.

Пример. Неопределенность доходности портфеля инвестиций

Представьте себе, что вы вложили $10000 в 200 акций некоторой корпорации, акции которой продаются по $50 за штуку. Ваш знакомый приобрел 100 акций этой же корпорации за $5000. Вы оба ожидаете, что стоимость акций возрастет в будущем году до $60 за акцию, что соответствует ставке прибыли 20%, (60 – 50) / 50. Оба вы также считаете маркетинговую стратегию этой корпорации довольно рискованной, поскольку она характеризуется стандартным отклонением курса акций $9. Это означает, что, хоть вы и ожидаете, что стоимость одной акции составит в будущем году $60, для вас не окажется неожиданным, если она будет примерно на $9 больше или меньше этого значения.

Вы предполагаете, что объем ваших инвестиций вырастет в будущем году до $12000 ($60 × 200), со стандартным отклонением $1800 ($9 × 200; более детальное описание приведено в разделе 5.4). Инвестиции вашего знакомого, как ожидается, в следующем году вырастут до $6000, со стандартным отклонением $900.

Складывается впечатление, что ваш риск (стандартное отклонение в $1800) в два раза больше, чем риск вашего знакомого ($900). И это действительно так, поскольку ваши инвестиции в абсолютном выражении в два раза больше. Однако оба вы делаете вложение в одни и те же ценные бумаги, а именно в акции одной и той же корпорации. Таким образом, во всех отношениях, за исключением объема инвестиций, ваша подверженность риску будет одинаковой. В относительном выражении (относительно объема первоначальных вложений) риски оказываются одинаковыми. В этом можно убедиться, вычислив коэффициент вариации (стандартное отклонение для стоимости акций в будущем году, деленное на среднее или ожидаемое значение). Коэффициент вариации в вашем случае будет равен $1800 / $12000 = 0,15, и он равен коэффициенту вариации для инвестиций вашего знакомого $ 900 / $6000 = 0,15. При этом и вы, и ваш знакомый будете считать, что неопределенность (или риск) составляет порядка 15% от ожидаемой в следующем году стоимости портфеля инвестиций.

Пример. Производительность труда в отделе торговли по телефону

Рассмотрим отдел торговли по телефону, в котором работают 19 сотрудников, занимающихся продажей билетов на концерт симфонической музыки. В среднем каждый сотрудник продает 23 билета в час. Стандартное отклонение составляет 6 билетов в час. Это означает, что любой из сотрудников может продавать в час в среднем на 6 билетов больше или меньше среднего значения (23 билета).

Если воспользоваться коэффициентом вариации и выразить различия в работе сотрудников в относительных величинах, можно легко вычислить, что эта величина составляет 6 / 23 = 0,261, или 26,1%. Это означает, что вариация производительности труда сотрудников составляет приблизительно 26,1% от среднего уровня продаж.

Для целей проведения анализа на высоком уровне и для формирования используемой менеджерами высшего звена стратегии число 26,1% (коэффициент вариации) может оказаться значительно полезнее, чем информация о различии в 6 билетов в час (стандартное отклонение). Менеджеры могут рассматривать отдельно уровень производительности труда (23 билета на одного сотрудника в час) и вариацию производительности (обычно производительность труда сотрудников отличается от средней не более чем на 26,1% в сторону больших или меньших значений).

Использование коэффициента вариации оказывается особенно полезным при проведении сравнений в условиях различных объемов. Рассмотрим еще один отдел торговли по телефону, занимающийся прода-

жей билетов в театры, в котором средний уровень продаж составляет 35 билетов в час, а стандартное отклонение равно 7. Поскольку производительность труда при продаже театральных билетов оказывается в целом выше производительности при продаже билетов на концерты симфонической музыки (средние значения соответственно 35 и 23), естественно, что и вариации здесь оказываются выше (7, а не 6). Однако коэффициент вариации для отдела, работающего с театральными билетами, составляет 7 / 35 = 0,200, или 20,0%. Сравнивая эту величину с коэффициентом 26,1%, характеризующим вариацию продаж билетов на симфонические концерты, менеджеры могут сделать вывод о том, что работающая с театральными билетами группа фактически более однородна (с точки зрения производительности отдельных сотрудников), чем группа, занятая продажей билетов на концерты симфонической музыки.

5.4. Результаты прибавления константы или изменения шкалы

Если ситуация изменяется определенным систематическим образом, то необходимость в пересчете таких характеристик, как типичные значения (среднее значение, медиана, мода), перцентили или меры изменчивости (стандартное отклонение, размах, коэффициент вариации) не возникает. Существует несколько основных правил быстрого вычисления соответствующих показателей для изменившейся ситуации.

Если к каждому значению данных *прибавляется* фиксированная величина, для получения соответствующих характеристик полученного таким образом нового набора данных эту же величину необходимо прибавить к среднему, медиане, моде и перцентилям исходного набора данных. Так, например, прибавление нового сбора в $5 к счетам, равным $38, $93, $25 и $89, означает, что эти счета окажутся равными $43, $98, $30 и $94. Средняя величина счета выросла ровно на $5, с $61,25 до $66,25. Вместо того чтобы рассчитывать среднее значение для новых счетов, можно просто прибавить $5 к найденному ранее среднему значению. Это правило применимо и для других типичных значений. Так, например, медиана в данном случае возрастает на $5, с $63,50 до $68,50. Однако стандартное отклонение и размах остаются прежними, поскольку сдвиг значений сохраняет между ними прежние расстояния. Коэффициент вариации изменяется, но его можно легко рассчитать, исходя из стандартного отклонения и среднего значения нового набора.

Если каждое значение данных *умножается* на фиксированное число, для получения среднего, медианы, моды, перцентилей, стандартного отклонения и размаха нового набора данных соответствующие показатели исходного набора необходимо умножить на это же число. Коэффициент вариации остается без изменений[15].

Если все входящие в набор данных величины умножаются на множитель c и к ним прибавляется величина d, приведенные выше правила действуют совместно: величина X превращается в $cX + d$. Новое среднее значение при этом оказывается равным $c \times$ (старое среднее) $+ d$. Аналогичные изменения претерпевают

[15] Здесь предполагается, что постоянный множитель больше нуля. В случае отрицательного множителя для вычисления стандартного отклонения и размаха необходимо использовать его абсолютное значение.

медиана, мода и перцентили. Новое стандартное отклонение равно $|c| \times$ (старое стандартное отклонение). Аналогичным образом корректируется и размах (обратите внимание на то, что величина прибавляемого значения d в этом случае не оказывает никакого влияния)[16]. Новый коэффициент вариации легко можно вычислить на основе новых значений среднего и стандартного отклонения.

В табл. 5.4.1 представлены описанные выше правила. Новый коэффициент вариации легко вычислить, воспользовавшись новыми значениями стандартного отклонения и среднего.

Таблица 5.4.1. Результаты прибавления константы или изменения шкалы

	Исходные данные	Прибавление d	Умножение на c	Умножение и прибавление
Величины данных	X	$X+d$	cX	$cX+d$
Среднее значение (то же для медианы, моды и перцентилей)	$\bar{X}$	$\bar{X}+d$	$c\bar{X}$	$c\bar{X}+d$
Стандартное отклонение (то же для размаха)	S	S	$\|c\|S$	$\|c\|S$

Пример. Неопределенность размера затрат в японских иенах и долларах США

Представьте себе, что расположенное за границей производственное подразделение вашей фирмы сформировало бюджет затрат на предстоящий год в следующем виде.

Ожидаемые затраты	325 700 000 японских иен
Стандартное отклонение	50 000 000 японских иен

Для составления общей финансовой сметы фирмы необходимо перевести эти значения в доллары США. Рассмотрим для простоты только коммерческий риск (представленный стандартным отклонением). Более подробный анализ может потребовать учета риска, связанного с возможными изменениями курсов обмена валют.

Японская иена легко переводится в доллары США с использованием текущего значения курса обмена валют[17]. Для перевода иены в доллар необходимо умножить известную величину предполагаемых затрат на число 0,007224, обозначающее, сколько долларов обменивается на одну иену. Умножая объем ожидаемых затрат и стандартное отклонение на этот переводной коэффициент, находим ожидаемые затраты и риск в долларовом выражении (величины округлены до тысяч).

Ожидаемые затраты	$2 353 000
Стандартное отклонение	$361 000

В данном примере использование приведенных нами основных правил дало возможность перевести данные, представленные в иенах, в долларовое выражение без повторного полного составления сметы в долларах.

[16] Стандартное отклонение умножается на абсолютную величину множителя, так что оно остается положительным. Например, при $c = -3$ стандартное отклонение умножается на 3.

[17] Его можно найти, например, в разделе "Обмен валют" *The Wall Street Journal*. В приведенном примере используется значение от 5 июня 1998 года.

Пример. *Общая стоимость произведенного товара*

При исчислении себестоимости и в финансовом деле производственные затраты часто разделяют на *фиксированные* затраты и *переменные* издержки на единицу продукции. Фиксированные затраты не зависят от количества произведенных единиц продукции, в то время как переменные издержки закладываются в смету для каждой единицы производимой продукции. Фиксированные затраты могут включать арендную плату и инвестиции в производственное оборудование, в то время как переменные издержки могут представлять стоимость реально использованных для производства материалов.

Рассмотрим производство шампуня, для которого фиксированные затраты составляют $1 000 000 в месяц, а переменные издержки равны $0,50 на один флакон. На основе тщательного анализа рыночного спроса менеджеры предусмотрели в следующем месяце выпуск 1 200 000 флаконов шампуня. Исходя из предыдущего опыта неопределенность для прогнозируемого объема производства можно оценить на уровне порядка 250000 флаконов. Таким образом, ожидается выпуск в среднем 1200000 флаконов шампуня, со стандартным отклонением 250 000 флаконов.

Если для объема производства существует такой прогноз, то каким будет прогноз для затрат? Обратите внимание на то, что объем производства переводится в затраты путем умножения количества единиц товара на $0,50 (переменные издержки) с прибавлением $1000000 (фиксированные затраты). Таким образом, в нашем случае

$$\text{общая стоимость} = \$0{,}50 \times 1\,200\,000 + \$1\,000\,000 = \$1\,600\,000,$$

$$\text{стандартное отклонение стоимости} = \$0{,}50 \times 250\,000 = \$125\,000$$

Итак, смета затрат составлена. Ожидаются затраты $1 600 000 со стандартным отклонением (неопределенностью) $125 000.

Коэффициент вариации для количества единиц произведенной продукции равен 250 000 / 1 200 000 = 20,8%. Коэффициент вариации для затрат также легко вычисляется, он равен $125 000 / $1 600 000 = 7,8%. Обратите внимание, что относительная вариация в стоимостном выражении оказывается значительно меньше, поскольку большие фиксированные затраты приводят к увеличению базы сравнения и соответственно к заметному снижению значения вариации.

5.5. Дополнительный материал

Резюме

Изменчивость (которую также называют **разнообразием, неопределенностью, рассеянием, разбросом** и **вариацией**) представляет собой меру различия отдельных значений набора данных между собой. В то время как величины, характеризующие центр (такие как среднее значение, медиана, мода) указывают типичную для набора данных *величину* значений, изменчивость показывает, *насколько близко* к этому центру обычно располагаются отдельные значения набора данных. Если все величины данных одинаковы, изменчивость равна нулю. Чем больше разброс величин, тем больше изменчивость.

Стандартное отклонение, которое обычно используют в качестве характеристики изменчивости, отражает типичное расстояние между средним значением и отдельными значениями набора данных. Стандартное отклонение показывает степень случайности в расположении отдельных значений относительно их общего среднего. **Отклонения** — это расстояния между каждым из значений и средним значением набора данных. Положительные отклонения соответствуют значениям, превышающим среднее, а отрицательные отклонения — значениям, меньшим среднего. Усреднение этих отклонений всегда дает результат, равный нулю. Стандартное отклонение показывает типичную величину таких отклонений (знак "минус" при этом не учитывается) и представляет собой число, измеряемое в тех же единицах, что и исходные данные (например, в долларах, в милях на один галлон или в килограммах).

Чтобы вычислить стандартное отклонение, необходимо выполнить следующее.

1. Найти отклонения, вычитая из каждого значения набора данных среднее.
2. Возвести полученные величины отклонений в квадрат, сложить их и разделить полученную сумму на n–1. Полученный результат называется *дисперсией*.
3. Извлечь квадратный корень. Полученное значение и есть стандартное отклонение.

При работе с данными обо всей генеральной совокупности необходимо использовать **стандартное отклонение генеральной совокупности** (обозначается буквой σ). В том случае, если необходимо сделать обобщение и перейти от имеющегося набора данных к некоторому большему множеству (реальному или гипотетическому), используется **стандартное отклонение выборки** (обозначается буквой *S*). При возникновении сомнений в том, какую их этих величин применить, нужно использовать стандартное отклонение выборки. Формулы для нахождения названных величин имеют следующий вид:

$$S = \sqrt{\frac{\text{Сумма квадратов отклонений}}{\text{Количество элементов в выборке} - 1}} = \sqrt{\frac{\left(X_1 - \bar{X}\right)^2 + \left(X_2 - \bar{X}\right)^2 + \ldots + \left(X_n - \bar{X}\right)^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2};$$

$$\sigma = \sqrt{\frac{\text{Сумма квадратов отклонений}}{\text{Количество элементов генеральной совокупности}}} = \sqrt{\frac{\left(X_1 - \mu\right)^2 + \left(X_2 - \mu\right)^2 + \ldots + \left(X_n - \mu\right)^2}{N}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2}.$$

В обеих этих формулах находят квадраты отклонений, делят их на соответствующую величину, а затем извлекают квадратный корень, чтобы от квадратов снова вернуться к исходным отклонениям. При вычислении стандартного отклонения выборки делят на n–1, поскольку отклонения вычисляют на основе неопределенного среднего значения выборки, а не на основе точного среднего значения генеральной совокупности.

Дисперсия — это квадрат стандартного отклонения. Эта величина несет ту же информацию, что и стандартное отклонение. Однако интерпретация дисперсии затруднена тем, что единицы измерения дисперсии представляют собой квадрат

единиц измерения исходных данных (например, доллар в квадрате, квадратные мили на один галлон в квадрате или килограммы в квадрате, независимо от содержательного смысла таких единиц). В связи с этим в качестве характеристики изменчивости чаще используют стандартное отклонение.

Если данные имеют нормальное распределение, стандартное отклонение равно приблизительно половине длины отрезка числовой прямой, который содержит две трети всех значений набора данных. Это означает, что приблизительно две трети всех значений находятся на расстоянии не более одной величины стандартного отклонения от среднего (выше или ниже среднего). Приблизительно 95% всех значений находятся на расстоянии не более двух величин стандартного отклонения от среднего, а около 99,7% значений лежат в пределах трех стандартных отклонений от среднего. Однако не следует ожидать справедливости этих утверждений для других (отличающихся от нормального) распределений.

Размах равен разности между максимальным и минимальным значениями набора данных. Эта величина характеризует протяженность, или ширину, набора данных. Размах используют как для описания данных, так и для поиска проблем в данных (в частности, для поиска ошибок при записи значений). Как статистическая характеристика размах имеет тот недостаток, что он акцентирует внимание только на экстремальных значениях и не учитывает типичные значения. Для большинства целей статистического анализа в качестве меры изменчивости более полезно использовать стандартное отклонение.

Коэффициент вариации равен частному от деления стандартного отклонения на среднее значение и характеризует относительную изменчивость данных, выраженную в долях или процентах от среднего. Коэффициент вариации — безразмерная величина. Он может быть полезен при сравнении изменчивости наборов данных, представленных в различных единицах измерения.

Прибавление фиксированного числа ко всем значениям набора данных приводит к увеличению среднего, медианы, перцентилей и моды на такое же число; стандартное отклонение и размах при этом не изменяются. При умножении каждого из значений набора данных на фиксированное число все характеристики — среднее, медиана, перцентили, мода, стандартное отклонение и размах — умножаются на это же число, а коэффициент вариации не изменяется[18]. При умножении каждого из значений данных на некоторое число и прибавлении другого фиксированного числа два описанных выше правила действуют совместно. Коэффициент вариации можно легко определить после того, как с применением этих правил вычисляется среднее и стандартное отклонения.

Основные термины

- Изменчивость (variability), разнообразие (diversity), неопределенность (uncertainly), рассеяние (dispersion), разброс (spread), 170
- Стандартное отклонение (standard deviation), 171
- Отклонение (deviation), 172
- Дисперсия (variance), 172
- Стандартное отклонение выборки (sample standard deviation), 187
- Стандартное отклонение генеральной совокупности (population standard deviation), 187
- Размах (range), 189
- Коэффициент вариации (coefficient of variation), 192

[18] Стандартное отклонение и размах умножаются на абсолютное значение этого числа и таким образом остаются положительными.

Контрольные вопросы

1. Что такое изменчивость?
2. а) Какие показатели обычно используют в качестве меры изменчивости?

   б) Какие еще показатели используют для этой цели?
3. а) Что такое отклонение от среднего значения?

   б) Чему равно среднее значение всех отклонений?
4. а) Что такое стандартное отклонение?

   б) Какую информацию о взаимосвязи между отдельными значениями и средним несет стандартное отклонение?

   в) В каких единицах измеряется стандартное отклонение?

   г) В чем состоит различие между стандартным отклонением выборки и стандартным отклонением генеральной совокупности?
5. а) Что такое дисперсия?

   б) В каких единицах измеряется дисперсия?

   в) Какую из мер изменчивости легче интерпретировать — стандартное отклонение или дисперсию? Почему?

   г) Если известно стандартное отклонение, дает ли дисперсия существенную дополнительную информацию об изменчивости?
6. Предположим, что некоторый набор данных имеет нормальное распределение. Какую часть значений можно в таком случае ожидать найти?

   а) На расстоянии не более одного стандартного отклонения от среднего значения?

   б) На расстоянии не более двух стандартных отклонений от среднего значения?

   в) В пределах трех величин стандартного отклонения от среднего значения?

   г) На расстоянии более одного стандартного отклонения от среднего?

   д) На расстоянии более одного стандартного отклонения *выше* среднего значения? (Будьте внимательны!)
7. Как изменятся ваши ответы на вопрос 6 в том случае, если данные не распределены нормально?
8. а) Что такое размах?

   б) В каких единицах измеряется размах?

в) В каких случаях используют такой показатель, как размах?

г) Полезен ли размах в качестве статистической меры изменчивости? Обоснуйте свой ответ.

9. а) Что такое коэффициент вариации?

   б) В каких единицах измеряется коэффициент вариации?

10. Какую меру изменчивости лучше использовать при сравнении изменчивости в двух разных ситуациях при условии, что средние в этих двух ситуациях сильно отличаются? Обоснуйте свой выбор.

11. Укажите, как изменятся в результате прибавления фиксированного числа к каждому значению следующие характеристики набора данных.

    а) Среднее, медиана, мода.

    б) Стандартное отклонение и размах.

    в) Коэффициент вариации.

12. Укажите, как изменятся при умножении на фиксированное число каждого значения следующие характеристики набора данных.

    а) Среднее, медиана, мода.

    б) Стандартное отклонение и размах.

    в) Коэффициент вариации.

Задачи

1. Представьте себе, что вы собираетесь создать рекламное агентство. Всегда существует много фирм, которые готовы сменить рекламное агентство, услугами которого они пользуются. В табл. 5.5.1 приведены размеры расходов на рекламу некоторых таких фирм.

   а) Определите средний объем затрат на рекламу.

Таблица 5.5.1. Расходы на рекламу

Фирма	Расходы на рекламу, млн дол.
Bell Atlantic Corp.	200
Chase Manhattan Corp.	40
Ernst & Young	15
L.A. Cellular	25
L.L. Bean	20
Merrill Lynch & Co.	70
Miramax Films	50
Pfizer	15
Qualcomm	20
Royal Caribbean and Celebrity Cruise Lines	40

Источник данных: *"Accounts in Play", Advertising Age*, May 11, 1998, p. 52.

б) Найдите стандартное отклонение объемов затрат, считая эту группу фирм выборкой из множества тех компаний, расходы на рекламу которых могут быть "захвачены" в любой момент времени. В каких единицах измеряется эта величина?

в) Дайте краткую интерпретацию стандартного отклонения (найденного в ответе на вопрос "б") для исследования различий между указанными фирмами.

г) Определите размах. В каких единицах он измеряется?

д) Дайте краткую интерпретацию размаха (найденного в ответе на вопрос "г") для исследования различий между рассматриваемыми фирмами.

е) Найдите коэффициент вариации. В каких единицах он измеряется?

ж) Дайте краткую интерпретацию коэффициента вариации (найденного в ответе на вопрос "е") для исследования различий между рассматриваемыми фирмами.

з) Вычислите дисперсию. В каких единицах она измеряется?

и) Дайте краткую интерпретацию дисперсии (найденной в ответе на вопрос "е") или поясните, почему простая интерпретация невозможна.

к) Постройте гистограмму для приведенного набора данных. Укажите на графике среднее значение, стандартное отклонение и размах.

2. Рассмотрим общую прибыль инвесторов, полученную за десять лет, с 1984 по 1994 гг., от ведущих аэрокосмических фирм. В табл. 5.2.2 приведены размеры прибыли, представленные как годовой уровень дохода в процентных единицах. У лидирующих фирм этой отрасли промышленности значение данного показателя может сильно различаться. Рассматриваемые фирмы можно считать выборкой, характеризующей деятельность крупных фирм всей аэрокосмической отрасти за соответствующий промежуток времени.

   а) Найдите стандартное отклонение общей прибыли. В каких единицах измеряется этот показатель?

   б) Дайте краткую интерпретацию стандартного отклонения (найденного в ответе на вопрос "а" для исследования различий между рассматриваемыми фирмами.

   в) Определите размах. В каких единицах он измеряется?

   г) Дайте краткую интерпретацию размаха (найденного в ответе на вопрос "в") для исследования различий между рассматриваемыми фирмами.

   д) Найдите коэффициент вариации. В каких единицах он измеряется?

   е) Дайте краткую интерпретацию коэффициента вариации (найденного в ответе на вопрос "д") для исследования различий между рассматриваемыми фирмами.

   ж) Вычислите дисперсию. В каких единицах она измеряется?

   з) Дайте краткую интерпретацию дисперсии (найденную в ответе на вопрос "ж") или поясните, почему эта величина не имеет простой интерпретации.

   и) Постройте гистограмму для приведенного набора данных. Укажите на графике среднее значение, стандартное отклонение и размах.

Таблица 5.5.2. Рентабельность фирм в аэрокосмической отрасли

Фирма	Общий доход инвесторов, %
Boeing	13
United Technologies	9
McDonnell Douglas	10
Lockheed	9
AlliedSignal	12
Martin Marietta	14
Textron	15
Northrop Grumman	6
General Dynamics	10
Gencorp	5
Sequa	0
Sundstrand	11
Thiokol	7
Rohr	–7
Kaman	5

Источник: *"The Fortune 1000 Ranked within Industries", Fortune*, May 15, 1995, p. F43– 44.

3. Исследуйте затраты на рекламу, приведенные в задаче 1, но выраженные не в долларах США, а во французских франках. Используйте величину обменного курса, приведенную в *The Wall Street Journal* или другом источнике. На основе ответов, которые вы получили при решении задачи 1 (т.е. без проведения новых расчетов), определите следующие значения.

   а) Среднюю величину затрат во французских франках.

   б) Стандартное отклонение.

   в) Размах.

   г) Коэффициент вариации.

4. В течение 12 месяцев, предшествовавших сентябрю 1995 года, прибыльность крупных взаимных фондов была высокой, поскольку хорошо работал и сам фондовый рынок. Однако в связи с различиями инвестиционных стратегий наблюдались существенные различия между фондами. Ставки прибыли за соответствующий период указаны в табл. 5.5.3.

   а) Найдите среднюю ставку прибыли для этих фондов.

   б) Найдите стандартное отклонение и дайте краткую интерпретацию.

   в) Сколько фондов не выходит за пределы одного стандартного отклонения от среднего? Сравните полученное число с тем, которое ожидалось бы в случае нормального распределения.

Таблица 5.5.3. Крупнейшие взаимные фонды: ставки прибыли (за 12 месяцев до сентября 1995 года) и активы (по состоянию на 31 августа 1995 года)

Фонд	Ставка прибыли, %	Активы, млн дол.
Fidelity Magellan Fund	37,6	51613 дол.
Investment Co. of America	22,5	23543
Washington Mutual Inv.	29,6	16091
Fidelity Puritan	14,4	14503
Vanguard Index 500	29,6	14109
Vanguard Windsor	26,3	13427
20th Century Ultra	41,8	13363
Fidelity Contrafund	32,8	13279
Income Fund of America	20,5	12560
Fidelity Growth & Income	24,5	12299
Janus Fund	25,8	11747
Vanguard Wellington Fund	23,9	11213
Fidelity Asset Manager	10,1	11065
Fidelity Equity – Income II	18,6	10807
Europacific Growth	8,7	10040
Vanguard Windsor II	28,3	1825
Fidelity Equity – Income	21,6	9314
Fidelity Adv. Growth Opp.	24,5	8785
New Perspective Fund	18,7	8584
Dean Witter Divid. Growth	23,4	8241
Growth Fund of America	29,3	7525
Putnam Growth & Income A	25,6	7518
Templeton Growth	14,6	6962
Templeton Foreign Funds	6,6	6932
Fidelity Blue Chip Growth	28,9	6733
Merrill Global Alloc. B	13,6	6623
Vanguard Wellesley Income	20,7	6587
AIM Equity Constelln.	41,5	6519
American Mutual	22,6	6379
T. Rowe Price Int'l Stock	3,9	6370
Templeton World	15,5	5869
Franklin Cust. Income	13,3	5720
Fidelity Growth Company	38,7	5534
20th Century Growth	30,9	5106

Окончание табл. 5.5.3

Фонд	Ставка прибыли, %	Активы, млн дол.
Pioneer II	19,9	5056
Putnam Voyager A	31,7	5051
Vanguard Institutional Index	29,8	5041
Fidelity Balanced	8,2	4981
Mutual Shares	25,3	4920
Fidelity Value Fund	18,6	4999

Данные взяты из *"Mutual Funds Zoom, but Stock Pros Predict a Slowdown", The Wall Street Journal*, October 5, 1995, p. R2. Источник данных: *Lipper Analytical Services*.

г) Сколько фондов находится в пределах двух стандартных отклонений от среднего? Сравните полученное число с тем, которое ожидалось бы в случае нормального распределения.

д) Сколько фондов находится в пределах трех стандартных отклонений от среднего? Сравните полученное число с тем, которое ожидалось бы в случае нормального распределения.

е) Постройте гистограмму для приведенного набора данных и укажите на графике пределы для одного, двух и трех стандартных отклонений от среднего значения. Поясните ответы на вопросы "в", "г" и "д" с позиций формы распределения.

5. Рассмотрите приведенные в табл. 5.5.3 активы взаимных фондов. Дайте ответы на вопросы предыдущей задачи, но не для ставки прибыли, а для активов.

6. Рассмотрим количество административных сотрудников во всех корпорациях Сиэтла с общим числом работников 500 и более[19]:



12, 15, 5, 16, 7, 18, 15, 12, 4, 3, 22, 4, 12, 4, 6, 8, 4, 5, 6, 4, 22, 10, 11, 4, 7, 6, 10, 10, 7, 8, 26, 9, 11, 41, 4, 16, 10, 11, 12, 8, 5, 9, 18, 6, 5.

а) Найдите среднее количество административных сотрудников в одной фирме.

б) Найдите стандартное отклонение (выборки) и дайте краткую интерпретацию этой величины.

в) Сколько фирм попадает в пределы одного стандартного отклонения от среднего? Сравните полученное количество с тем, которое ожидалось бы в случае нормального распределения.

г) Сколько фирм попадает в пределы двух стандартных отклонений от среднего? Сравните полученное количество с тем, которое ожидалось бы в случае нормального распределения.

[19] Данные взяты из "Pacific Northwest Executive", April 1988, p. 20.

д) Сколько фирм попадает в пределы трех стандартных отклонений от среднего? Сравните полученное количество с тем, которое ожидалось бы в случае нормального распределения.

е) Постройте гистограмму для приведенного набора данных и укажите на графике пределы для одного, двух и трех стандартных отклонений от среднего значения. Поясните ответы на вопросы "в", "г" и "д" с позиций формы распределения.

7. Получите ответы на вопросы задачи 6, опустив данные той фирмы, которая имеет максимальное отличие от среднего. Дайте краткое письменное (один абзац) сравнение результатов, полученных с учетом и без учета этой фирмы.

8. Все 18 сотрудников некоторого отдела получили 4-процентную прибавку к зарплате. Как изменились при этом следующие показатели?

а) Средняя зарплата по отделу?

б) Стандартное отклонение зарплат?

в) Размах зарплаты?

г) Коэффициент вариации зарплат?

9. На основе прогнозного анализа спроса предприятие планирует выпустить в текущем квартале в среднем 80 000 игровых видеокассет. Неопределенность оценивается в 25 000 кассет, что является стандартным отклонением объема выпуска. Фиксированные затраты для используемого оборудования равны $72 000 в квартал, а переменные издержки составляют $0,68 на одну произведенную кассету.

а) Чему равна прогнозируемая ожидаемая общая стоимость произведенных кассет?

б) Какова неопределенность в прогнозе общей стоимости продукции, если эту неопределенность выразить в виде стандартного отклонения?

в) Найдите коэффициенты вариации для количества произведенных кассет и для общей стоимости продукции. Сравните в письменном виде (один абзац) полученные коэффициенты вариации.

г) В конце квартала оказалось, что предприятие выпустило 10000 кассет. На сколько стандартных отклонений в сторону превышения или в сторону уменьшения отличается от среднего это количество?

д) Предположим, что фирма произвела 20000 кассет. На сколько стандартных отклонений в сторону превышения или в сторону уменьшения отличается от среднего это количество? Будет ли такой результат неожиданным в свете предыдущего прогноза? Обоснуйте свой ответ.

10. Представьте себе, что составлен рекламный бюджет вашей фирмы на следующий год. Вы (как менеджер отдела маркетинга) собираетесь потратить примерно $1 500 000 на услуги коммерческого телевидения. Неопределенность, которая характеризуется стандартным отклонением, составляет $200 000. Рекламное агентство, с которым вы работаете, берет за свои услуги 15% от этой суммы. Определите ожидаемый объем затрат на услуги рекламного агентства для своей фирмы и неопределенность этой величины.

11. Вы стараетесь контролировать вес шоколадного батончика с арахисовой начинкой, вмешиваясь для этой цели в процесс производства. В табл. 5.5.4 показан вес батончиков в двух выборках образцов из дневного производства. Одна выборка взята до вашего вмешательства, а другая — после него.

    а) Найдите средний вес шоколадных батончиков до вмешательства.

    б) Найдите стандартное отклонение веса до вмешательства.

    в) Найдите средний вес шоколадных батончиков после вмешательства.

    г) Найдите стандартное отклонение веса после вмешательства.

    д) Сравните стандартные отклонения до и после вашего вмешательства. Опишите (один абзац) достигнутый результат и, в частности, то, удалось ли вам снизить изменчивость результатов данного производственного процесса?

12. Всем нам приходилось время от времени испытывать неудобства, связанные с транспортными пробками, в результате которых машины еле движутся по автостраде. Если у вас есть пассажир (или сотовый телефон), вы легче сможете справиться с этой проблемой, но во что транспортные заторы обходятся обществу в целом? Обратимся к данным, приведенным в табл. 5.5.5.

    а) Охарактеризуйте с помощью среднего потери от транспортных заторов во всех указанных в таблице городах.

    б) Охарактеризуйте с помощью стандартного отклонения изменчивость потерь от транспортных пробок в различных городах. Приведенный набор данных можно рассматривать в качестве генеральной совокупности, содержащей всю доступную информацию.

    в) Постройте гистограмму для приведенного набора данных. Покажите на графике среднее значение и стандартное отклонение.

    г) Кратко опишите (один абзац), что вам удалось выяснить при исследовании приведенных данных.

Таблица 5.5.4. Вес двух выборок шоколадных батончиков (в унциях)

До вмешательства	До вмешательства	После вмешательства	После вмешательства
1,62	1,68	1,60	1,69
1,71	1,66	1,71	1,59
1,63	1,64	1,65	1,66
1,62	1,70	1,64	1,68
1,63	1,66	1,63	1,59
1,69	1,71	1,65	1,57
1,64	1,63	1,74	1,62
1,63	1,65	1,75	1,75
1,62	1,70	1,66	1,72
1,70	1,74	1,73	1,63

Таблица 5.5.5. Потери от транспортных заторов в расчете на одно зарегистрированное транспортное средство

Города северо-восточного региона США					
Балтимор	530	Нью-Йорк	1090	Питсбург	400
Бостон	880	Филадельфия	420	Вашингтон	1420
Хартфорд	250				
Города региона Среднего Запада США					
Чикаго	570	Детройт	530	Милуоки	370
Цинциннати	200	Индианаполис	130	Миннеаполис Сент-Пол	270
Кливленд	140	Канзас-Сити	160	Оклахома-Сити	190
Колумбус	230	Луисвилл	190	Сент-Луис	540
Города южного региона США					
Атланта	640	Мемфис	140	Норфолк	390
Шарлотт	390	Майами	680	Орландо	420
Форт-Лодердейл	290	Нашвилл	340	Тампа	310
Джэксонвилл	400	Новый Орлеан	340		
Города юго-западного региона США					
Альбукерк	210	Денвер	420	Финикс	630
Остин	410	Эль-Пасо	120	Солт-Лейк-Сити	90
Корпус-Кристи	50	Форт-Уэрт	420	Сан-Антонио	290
Даллас	750	Хьюстон	750		
Города западного региона США					
Гонолулу	470	Сакраменто	280	Сан-Франциско – Окленд	930
Лос-Анджелес	980	Сан-Бернардино-Ривер	1320	Сан-Хосе	960
Портленд	500	Сан-Диего	480	Сиэтл – Еверет	880

Источник данных: *U.S. Bureau of the Census, Statistical Abstract of the United States: 1994, 11th ed. (Washington D.C., 1994), Table 1013.* При построении таблицы использованы различные источники информации федерального уровня, отдельных штатов и местные источники. Основной источник информации: *Federal Highway Administration's Highway Performance Monitoring System. Texas Transportation Institute, College Station, Texas; Roadway Congestion in Major Urban Areas*, ежегодное издание (авторские права защищены).

13. Ниже приведены ставки дохода для выборки выполненных в последнее время контрактов на внутрисистемное обслуживание:

    78,9%; 22,5%; –5,2%; 997,3%; –20,7%; –13,5%; 429,7%; 88,4%; –52,1%; 960,1%; –38,8%; –70,9%; –73,3%; 47,0%; –1,5%; 23,9%; –35,6%; –62,0%; –75,7%; –14,0%; –81,2%; 46,9%; 135,1%; –34,6%; –85,3%; –73,6%; –9,0%; 19,6%; –86,7%; –87,6%; –88,7%; –75,5%; –91,0%; –97,9%; –100,0%

а) Найдите среднее значение и стандартное отклонение для этих ставок дохода.

б) Опишите (в одном абзаце) уровень риска (затраты) и средний доход (прибыль) в данной отрасли.

14. Рассмотрим процентные ставки начислений по счетам для некоторой выборки местных банков:

3,00%; 4,50%; 4,90%; 3,50%; 4,75%; 3,50%; 3,50%; 4,25%; 3,75%; 4,00%

а) Найдите стандартное отклонение процентных ставок.

б) Что можно сказать о банках данного региона исходя из полученной величины стандартного отклонения?

15. Рассмотрим процентные колебания курса доллара относительно других валют за четыре недели (табл. 5.5.6).

а) Найдите стандартное отклонение для приведенных величин.

б) Дайте интерпретацию вычисленного стандартного отклонения. В частности, что этот показатель означает для валютных рынков?

16. Ниже приведен вес произведенных за последнее время раковин:

20,8; 20,9; 19,5; 20,8; 20,0; 19,8; 20,1; 20,5; 19,8; 20,3; 20,0; 19,7; 20,3; 19,5; 20,2; 20,2; 19,5; 20,5.

Вычислите обычно используемую меру, приблизительно характеризующую отличие этих весов от среднего.

17. Рассмотрите стоимость обеда для двух человек (включая бутылку вина, налоги и стоимость обслуживания) в шести европейских городах (табл. 5.5.7).

а) Найдите на основе приведенных данных среднюю стоимость обеда на двоих в Европе.

б) Найдите стандартное отклонение выборки для этих цен.

в) Что можно сказать о стоимости обеда в Европе на основе полученного значения стандартного отклонения?

18. Рассмотрим стоимость (в тыс. дол.) подарков, которые были возвращены в каждый из универмагов компании по окончании праздников (табл. 5.5.8).

Таблица 5.5.6. Колебание курса доллара

Страна	Колебания, %	Страна	Колебания, %
Бельгия	–5,3	Сингапур	–1,5
Япония	–6,7	Франция	–4,9
Бразилия	26,0	Южная Корея	–1,0
Мексика	–1,2	Гонконг	0,0
Великобритания	–3,7	Тайвань	–0,1
Нидерланды	–5,1	Италия	–4,7
Канада	–1,9	Германия	–5,1

Таблица 5.5.7. Стоимость обеда

Город	Цена, дол.
Лондон	72,42
Париж	79,30
Рим	64,77
Мадрид	66,26
Брюссель	61,15
Вена	78,10

Данные взяты из раздела путешествий и экскурсий, *New York Times*, October 18, 1987, p. 14.

Таблица 5.5.8. Возвращенные подарки

Магазин	Возврат, тыс. дол.
A	13
B	8
C	36
D	18
E	6
F	21

а) Вычислите стандартное отклонение выборки.

б) Опишите (один абзац) интерпретацию стандартного отклонения для характеристики различий между магазинами.

19. Билеты на самолеты обычно предлагают, руководствуясь выгодой для перевозчика, а не для пассажира. 21 июля 1998 года работающее с использованием компьютерных сетей туристическое агентство *Expedia* предлагало следующие тарифы для перелета из Сиэтла в Бостон и обратно на сентябрь 1998 года: *United Airlines*: $708,00; *America West*: $930,00; *Northwest Airlines*: $1504,50; *U.S. Airways*: $1795,50; *Delta Air Lines*: $1798,50.

а) Вычислите стандартное отклонение, считая эти тарифы выборкой из множества тарифов, которые могут быть предложены в подобных обстоятельствах.

б) Опишите (один абзац) интерпретацию вычисленного стандартного отклонения и обсудите вопрос о различиях тарифов, предлагаемых разными авиалиниями.

20. Рассмотрим следующие значения показателя производительности труда для генеральной совокупности работников:

85,7; 78,1; 69,1; 73,3; 86,8; 72,4; 67,5; 76,8; 80,2; 70,0.

а) Вычислите среднюю производительность.

б) Вычислите и дайте интерпретацию стандартного отклонения производительности.

в) Вычислите и дайте интерпретацию коэффициента вариации производительности.

г) Вычислите и дайте интерпретацию размаха значений производительности.

21. Ниже приведены результаты продаж (в тыс.) за первый год для некоторых выпущенных на рынок новых товаров, аналогичных тому, который вы рассматриваете.

10, 12, 16, 47, 39, 22, 110, 29

а) Вычислите среднее значение и стандартное отклонение. Дайте интерпретацию стандартного отклонения.

б) Для нового товара, который вы рассматриваете, объем продаж за первый год составил 38 тыс. Насколько это значение отличается от среднего? Сравните свой ответ с тем, что следует из величины стандартного отклонения.

в) Для следующего нового вида товара объем продаж в первый год после выпуска на рынок составил 92 тыс. Насколько в величинах стандартного отклонения (найденного в ответе на вопрос "а") это значение отличается от среднего (найденного в ответе на вопрос "а")?

г) Для каждого из рассматриваемых случаев выпуска нового товара (вопросы "б" и "в") укажите, типичен ли данный случай для вашей фирмы, и обоснуйте свой вывод.

22. Взятые из шахты образцы показали следующее процентное содержание золота:

1,1; 0,3; 1,5; 0,4; 0,8; 2,2; 0,7; 1,4; 0,2; 4,5; 0,2; 0,8.

а) Вычислите и дайте интерпретацию стандартного отклонения для данной выборки.

б) Вычислите и дайте интерпретацию коэффициента вариации.

в) Для какого из значений наблюдается максимальное положительное отклонение? Почему место, из которого взят этот образец, представляет особый интерес?

г) На сколько величин стандартного отклонения превышает среднее то значение, которое имеет максимальное положительное отклонение?

23. Рассмотрим доходность акций компаний из некоторой выборки, представленную в виде годовой процентной ставки:

5,5; 10,6; 19,0; 54,5; 6,6; 26,8; 6,2; −2,4; −28,3; 2,3.

а) Вычислите среднее значение и стандартное отклонение доходности акций.

б) Дайте интерпретацию стандартного отклонения.

в) Насколько ниже среднего уровня (в величинах стандартного отклонения) оказывается самая низкая доходность?

г) Постройте гистограмму и укажите на ней среднее значение, стандартное отклонение, а также отклонение акций с самой низкой доходностью.

24. Предполагаемые затраты фирмы в среднем равны $138 000, а стандартное отклонение составляет $35 000. Как оказалось, поставщики фирмы повышают цены в целом на 4%. Как изменятся среднее значение и стандартное отклонение затрат?

25. Для предыдущей задачи сравните коэффициенты вариации до и после повышения цен. Почему изменяется (или не изменяется) этот показатель?

26. Представьте себе, что вы работаете менеджером по продажам в региональном отделении компании, производящей напитки. Планы продаж для ваших представителей имеют среднее значение $768 000, а стандартное отклонение составляет $240 000. Вы получили распоряжение повысить планы продаж для каждого из торговых представителей на $85 000. Как при этом изменится стандартное отклонение?

27. Для данных предыдущей задачи сравните коэффициенты вариации до и после изменения планов продаж. Почему изменяется (или не изменяется) этот показатель?

28. Найдите стандартное отклонение размеров налога на добавленную стоимость, приведенных в табл. 4.3.2. Что можно сказать на основе этих данных о практике налогообложения в разных странах?

29. Рассмотрите данные о длительности показа фильмов, приведенные в табл. 4.3.7.

а) Вычислите стандартное отклонение. Как это значение характеризует продолжительность рассматриваемых фильмов?

б) Определите размах. Как он характеризует продолжительность рассматриваемых фильмов?

в) На сколько величин стандартного отклонения отличается от среднего значения время показа самого длинного фильма?

30. В разных странах используются разные стратегии налогообложения. В некоторых случаях увеличивают налог на прибыль, в других основное внимание обращается на налоги на товары и услуги. Рассмотрим относительную величину налогов на товары и услуги, предусмотренных в ставках налогов некоторых стран (табл. 5.5.9). Эти относительные величины определяются двумя способами: как сумма налогов на все товары и услуги, выраженная в процентах от общего объема экономической деятельности (валовой внутренний продукт), и в процентах от всех собранных в стране налогов.

а) Найдите стандартное отклонение для каждой из переменных.

б) Найдите размах для каждой из переменных.

в) Найдите коэффициент вариации для каждой из переменных.

г) Сравните полученные характеристики изменчивости. Для какой из характеристик изменчивости при переходе от одной переменной к другой наблюдаются наименьшие различия? Почему?

Таблица 5.5.9. Международное налогообложение: налоги на товары и услуги в процентном отношении к валовому внутреннему продукту (ВВП) и налоговым поступлениям

	Налоги на товары и услуги, в % от			Налоги на товары и услуги, в % от	
Страна	ВВП	Налоги	Страна	ВВП	Налоги
Бельгия	7,2	16,0	Нидерланды	7,3	15,6
Канада	5,3	14,1	Новая Зеландия	8,6	23,8
Дания	9,9	20,6	Норвегия	8,2	17,4
Франция	7,9	17,8	Португалия	6,8	19,0
Германия	6,4	16,4	Испания	5,5	15,9
Греция	9,9	25,9	Швейцария	3,0	9,7
Италия	5,7	14,3	Турция	6,5	22,2
Япония	1,4	4,4	Великобритания	6,7	18,5
Люксембург	7,2	14,9			

Данные взяты из статьи *Gilbert E. Metcalf, "Value-Added Taxation: A Tax Whose Time Has Come?" Journal of Economic Perspectives 9, No. 1* (Winter 1995), p. 129. Источник данных: *"Price Waterhouse Guide to Doing Business in ..."* для различных стран, *Eurostat (1993)* и *OECD Revenue Statistics*.

31. Для данных о налогообложении товаров и услуг, приведенных в табл. 5.5.9:

а) Постройте для каждой из переменных в одинаковом масштабе блочные диаграммы.

б) Прокомментируйте, используя эти диаграммы, распределение каждой из переменных.

32. Для данных о доходности не подлежащих налогообложению облигаций (задача 6 из главы 3):

а) Найдите стандартное отклонение доходности.

б) Найдите размах.

в) Найдите коэффициент вариации.

г) Воспользуйтесь полученными значениями для описания уровня изменчивости доходности.

33. На основе данных задачи 7 из главы 3 найдите стандартное отклонение и размах, характеризующие типичную изменчивость (или неопределенность) реакции рынка на объявления о выкупе компанией собственных акций.

34. На основе данных задачи 9 из главы 3, в которой рассматриваются портфельные инвестиции *CREF* в компании по производству мебели:

а) Найдите стандартное отклонение для рыночных цен акций этих фирм в портфеле *CREF*.

б) Найдите размах рыночных цен.

в) Найдите коэффициент вариации.

г) Охарактеризуйте изменчивость, воспользовавшись для этого тремя полученными значениями.

д) Сколько значений рыночных цен находится в пределах одной величины стандартного отклонения от среднего? Как это соотносится с тем количеством, которое можно было бы ожидать в случае нормального распределения?

е) Сколько значений рыночных цен находится в пределах двух величин стандартного отклонения от среднего? Как это соотносится с тем количеством, которое можно было бы ожидать в случае нормального распределения?

35. На основе данных задачи 16 из главы 3, используя стандартное отклонение, кратко охарактеризуйте изменчивость стоимости традиционных похоронных услуг.

36. Обратитесь к данным задачи 19 из главы 3 о низком качестве производства электромоторов.

а) Найдите стандартное отклонение и размах, чтобы охарактеризовать типичную изменчивость качества в партиях продукции.

б) Повторно вычислите стандартное отклонение и размах, не рассматривая два крайних значения.

в) Сравните величины стандартного отклонения и размаха с учетом и без учета этих крайних значений. В частности, дайте ответ на вопрос: насколько чувствительны обе эти характеристики к исключению из рассмотрения крайних значений?

37. Вычислите стандартное отклонение для данных задачи 2 из главы 4, чтобы определить изменчивость уровня затрат постоянных заказчиков за последний месяц. Опишите найденные различия в одном абзаце.

38. Насколько различаются затраты на научно-исследовательскую работу для разных фирм, работающих в пищевой промышленности? Найдите и прокомментируйте стандартное отклонение и размах для данных задачи 4 из главы 4.

39. Какова изменчивость платы за ссуды под залог недвижимости? Найдите и прокомментируйте стандартное отклонение, размах и коэффициент вариации для данных, приведенных в задаче 13 из главы 4.

40. Взаимные фонды часто заявляют более высокий уровень доходности, чем та, которую можно получить при реальном вложении в них денег. В табл. 5.5.10 показана годовая доходность для диверсифицированных международных фондов облигаций до и после вычета различных затрат, комиссионных выплат, сборов с продаж и подлежащих выплате налогов.

а) Найдите стандартное отклонение доходов до и после корректировки.

б) Такая корректировка увеличивает или уменьшает однородность этих фондов (рассматриваемых как группа)? Обоснуйте свой вывод.

41. Рассмотрим возраст (в годах) и эксплуатационные затраты (в тыс. дол. в год) для пяти одинаковых печатных машин (табл. 5.5.11).

Таблица 5.5.10. Доходность международных взаимных фондов облигаций

Фонд	Годовая доходность, %	
	до корректировки	без комиссионных и налогов
T. Rowe Price International Bond	6,3	3,3
Merrill Lynch Global Bond B	9,5	2,6
Merrill Lynch Global Bond A	10,4	2,1
IDS Global Bond	13,1	4,3
Merrill Lynch World Income A	6,5	–0,6
Fidelity Global Bond	4,6	2,3
Putnam Global Governmental Income	10,3	0,5
Shearson Global Bond B	7,5	1,2
Paine Webber Global Income B	3,4	–2,7
MFS Worldwide Governments	5,4	–2,5

Источник данных: *Fortune,* March 22, 1993, p. 156.

Таблица 5.5.11. Зависимость эксплуатационных затрат от срока эксплуатации для одинаковых печатных машин

Износ	Эксплуатационные затраты
2	6
5	13
9	23
3	5
8	22

а) Вычислите средний возраст (срок эксплуатации) печатных машин.

б) Вычислите стандартное отклонение возраста печатных машин.

в) Вычислите размах возраста печатных машин.

г) Вычислите коэффициент вариации возраста печатных машин.

42. Используя данные предыдущей задачи, в которой рассматриваются срок эксплуатации и эксплуатационные затраты для пяти одинаковых печатных машин:

а) Вычислите средние эксплуатационные затраты.

б) Вычислите стандартное отклонение эксплуатационных затрат.

в) Вычислите размах эксплуатационных затрат.

г) Вычислите коэффициент вариации эксплуатационных затрат.

43. Для изменений процентной ставки 20 компаний, значащихся в списке транспортного индекса Доу Джонс (*Dow Jones Transportation Average*) фондовой биржи за 31 августа 1998 года (задача 24 из главы 2):

а) Найдите стандартное отклонение изменения процентной ставки.

б) Определите размах изменения процентной ставки.

в) Найдите коэффициент вариации изменения процентной ставки.

44. Для значений транспортного индекса Доу Джонс (*Dow Jones Transportation Average*) за сентябрь 1998 (задача 25 из главы 2):

а) Найдите стандартное отклонение изменения индекса.

б) Найдите размах изменения.

в) Найдите коэффициент вариации изменения.

г) Найдите стандартное отклонение процентного изменения.

д) Определите размах процентного изменения.

е) Найдите коэффициент вариации процентного изменения.

Упражнения с использованием базы данных

Обратитесь к базе данных о наемных работниках в приложении А.

1. Для размера заработной платы за год:

а) Найдите размах.

б) Найдите стандартное отклонение.

в) Найдите коэффициент вариации.

г) Сравните эти три показателя. Как они характеризуют типичную заработную плату в рассматриваемом отделе?

2. Для размера заработной платы за год:

а) Постройте гистограмму и покажите на ней среднее значение и стандартное отклонение.

б) Сколько работников имеют зарплату, отличающуюся от средней не более чем на одну величину стандартного отклонения? Как это количество согласуется с тем числом, которое можно было бы ожидать в случае нормального распределения?

в) Сколько работников имеют зарплату, отличающуюся от средней не более чем на два стандартных отклонения? Как это количество согласуется с тем числом, которое можно было бы ожидать в случае нормального распределения?

г) Сколько работников имеют зарплату, отличающуюся от средней не более чем на три стандартных отклонения? Как это количество согласуется с тем числом, которое можно было бы ожидать в случае нормального распределения?

3. Для возраста сотрудников дайте ответы на вопросы упражнения 1.
4. Для возраста сотрудников дайте ответы на вопросы упражнения 2.
5. Для квалификации (опыта работы) сотрудников дайте ответы на вопросы упражнения 1.
6. Для квалификации (опыта работы) сотрудников дайте ответы на вопросы упражнения 2.

Проекты

1. В соответствии со своими интересами возьмите набор значений некоторого свойства (по вашему выбору), измеренного для предприятий из двух отраслей промышленности (не менее 15 предприятий в каждой группе).

   а) Для каждой группы:

   1) Охарактеризуйте изменчивость изучаемого свойства, воспользовавшись описанными в этой главе методами, которые могут быть применены к вашим данных.

   2) Для каждого из наборов данных изобразите полученные характеристики изменчивости на гистограмме и/или блочной диаграмме.

   3) Опишите (один абзац), что вы узнали о соответствующей отрасли промышленности на основе анализа изменчивости изучаемого свойства.

   б) Проведите для обеих групп следующие сравнения:

   1) Сравните величины стандартных отклонений.

   2) Сравните коэффициенты вариации.

   3) Сравните величины размаха.

   4) Кратко опишите, что вы узнали в результате сравнительного анализа изменчивости для рассматриваемых отраслей промышленности. В частности, какая из характеристик изменчивости оказалась наиболее полезной?

2. Возьмите набор данных, включающий не менее 25 значений, характеризующих интересующее вас предприятие или отрасль промышленности. Опишите эти данные, воспользовавшись всеми изученными к этому моменту методами, которые применимы к вашим данным. Используйте как численные, так и графические методы; обращайте внимание как на типичное значение, так и на изменчивость. Представьте полученные результаты в виде двухстраничного отчета для руководства, сформулировав рекомендации в первом абзаце.

Ситуация для анализа

Следует ли продолжать работу с этим поставщиком?

Вы и один из ваших сотрудников, Б.У. Келлерман, получили задание — оценить нового поставщика деталей к выпускаемому вашей фирмой оборудованию для ухода за домом и садом. Одна из деталей должна иметь размер 8,5 см, однако допускается также любой размер в пределах от 8,4 до 8,6 см. Келлерман недавно доложил об исследовании размеров 99 поставленных деталей. Сделанный Келлерманом первый набросок отчета содержит такие рекомендации.

Качество деталей, поставляемых фирмой HypoTech, не соответствует нашим требованиям. Несмотря на то что цены этой фирмы достаточно низкие и привлекательные, а поставки производятся в соответствии с графиком, качество изделий недостаточно высокое. Мы рекомендуем серьезно рассмотреть вопрос об использовании альтернативных источников поставок.

Теперь ваша очередь. После анализа полученных Келлерманом цифр и чтения проекта отчета перед вами стоит задача подтвердить его рекомендации (или отказаться от них) на основе собственного независимого исследования.

Выводы Келлермана представляются осмысленными. Основной аргумент состоит в том, что, несмотря на среднее значение, составляющее 8,494 см и очень близкое к стандарту в 8,5 см, стандартное отклонение довольно велико, оно составляет 0,103, в результате чего дефектные детали составляют примерно треть всех поставляемых изделий. Действительно, Келлерман явно горд тем, что помнит сведения, полученные давным-давно при изучении статистики, — что-то о том, что попадание в пределы одного стандартного отклонения от среднего наблюдается примерно в трети случаев. В данном конкретном случае при такой цене можно допустить 10, или даже 20% дефектных деталей, однако 30 или 33% выходит за рамки разумного.

Ситуация представляется совершенно очевидной, однако для того, чтобы убедиться в правильности полученных Келлерманом выводов, вы решаете все-таки самостоятельно быстро просмотреть данные. Естественно, вы ожидаете, что выводы подтвердятся. Вот этот набор данных:

8,503	8,503	8,500	8,496	8,500	8,503	8,497	8,504	8,503	8,506
8,502	8,501	8,489	8,499	8,492	8,497	8,506	8,502	8,505	8,489
8,505	8,499	8,89	8,505	8,504	8,499	8,499	8,506	8,493	8,494
8,510	8,310	8,804	8,503	8,782	8,502	8,509	8,499	8,498	8,493
8,346	8,499	8,505	8,509	8,499	8,503	8,494	8,511	8,501	8,497
8,501	8,502	8,780	8,494	8,500	8,498	8,500	8,502	8,501	8,491
8,511	8,494	8,374	8,492	8,497	8,150	8,496	8,501	8,489	8,506
8,493	8,498	8,505	8,490	8,493	8,501	8,497	8,501	8 498	8,503
8,508	8,501	8,499	8,504	8,505	8,461	8,497	8,495	8 504	8,501
8,493	8,504	8,897	8,505	8,490	8,492	8,503	8,507	8 497	

Вопросы для обсуждения

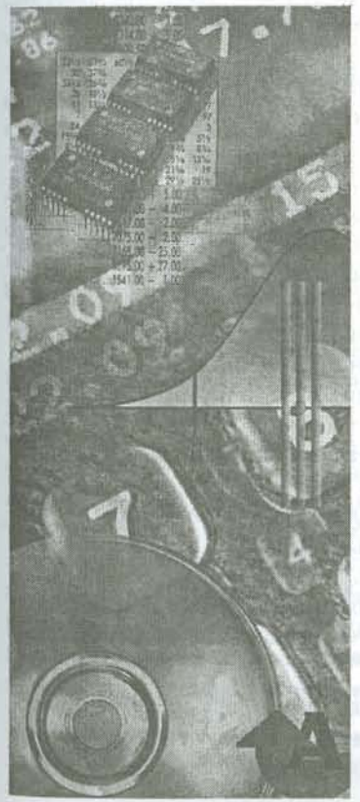
1. Правильны ли результаты вычислений Келлермана? Это первое, что необходимо проверить.
2. Внимательно посмотрите на данные, используя подходящие статистические методы.
3. Верны ли выводы, которые сделал Келлерман? Если да, почему вы так считаете? Если нет, то почему нет, и что следует сделать для выработки правильных рекомендаций?

Вероятность

В этой части...

Глава 6. "Вероятность: разбираемся в случайных ситуациях"
Глава 7. "Случайные величины: работа с неопределенными значениями"

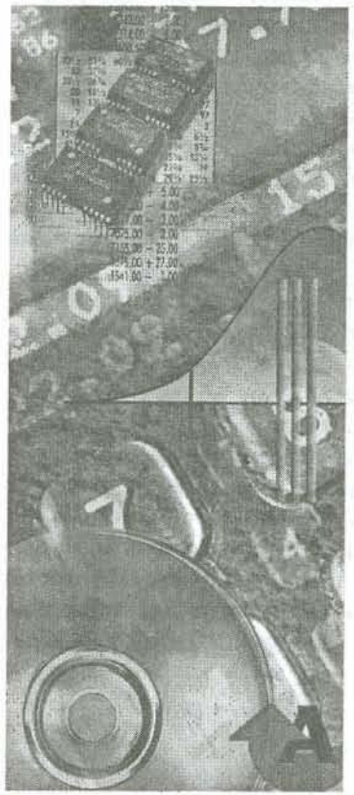
Как работать с неопределенностью? На основе понимания ее механизмов. Изучение *вероятности* вдохновляет нас на дальнейшее углубление своих представлений о неопределенности с целью научиться отличать действительно неопределенные процессы от того, в чем можно быть полностью уверенным. Что в действительности представляет собой интересующая нас неопределенная ситуация? Какой именно четкой процедурой она определяется? Насколько вероятны различные возможности? Часто приходится иметь дело с событиями, которые могут либо произойти, либо не произойти: удастся ли заключить контракт, поступит ли заявка от клиента, исправят ли вовремя оборудование? В главе 6 рассмотрены ситуации такого типа, их возможные комбинации, а также вопрос уменьшения неопределенности за счет получения дополнительной информации. Существуют также ситуации, в которых интерес представляют неопределенные *числа*, какие-то значения, которые точно неизвестны: каковы будут объявленные доходы, какими окажутся квартальные уровни продаж, сколько времени будет потеряно в результате возникшей поломки компьютера? В главе 7 мы рассмотрим методы нахождения соответствующего характерного числа, узнаем, как оценить риск (или неустойчивость) некоторой ситуации и определить правдоподобие различных сценариев, базирующихся на неопределенных числах.



ГЛАВА 6

Вероятность: разбираемся в случайных ситуациях

Нам необходимо уметь разбираться в случайных ситуациях, по крайней мере настолько, насколько это возможно. К сожалению, видимо, мы никогда не сможем "совершенно точно" сказать, что случится в будущем. Однако осознание того, что одни возможности представляются более правдоподобными, чем другие, и количественное описание (с помощью соответствующих чисел) этих отношений дает нам преимущества по сравнению с теми, кто вовсе не имеет представления о том, что произойдет, или полагается лишь на свои ощущения, не имея для этого объективных предпосылок. Лучше всего объединять понимание вероятностей со всеми доступными знаниями и опытом.



Ниже описано несколько ситуаций, в которых присутствует элемент неопределенности.

Случай первый. В вашем отделе должно быть принято решение о том, следует ли выводить на потребительский рынок новый цифровой аудиоплейер. Несмотря на то что проведенное маркетинговое исследование показало, что типичным потребителям он нравится и цена представляется им разумной, успех этого товара вряд ли можно считать гарантированным. Неопределенность существует вследствие ряда причин. Как, например, скажется конкуренция? Будут ли ваши поставщики своевременно предоставлять качественные компоненты? Не существует ли препятствий, которые пока еще не замечены? Будет ли в экономике наблюдаться спад или подъем? Будут ли потребители действительно платить за этот товар деньги, или они только заявляют об этом? Вашей фирме необходимо принять наилучшее решение на основе доступной информации.

Случай второй. Ваш дядя, фермер из Миннесоты, написал вам о том, что в этом году возможна засуха. Поскольку он правильно предсказал возникновение аналогичных проблем в 1988 году, вы сразу же приобретаете на чикагской торговой бирже опционы "колл" на зерно и заключаете фьючерсные контракты на сою. Что произойдет с вложенными при этом средствами? Ответить на этот вопрос очень непросто. Если урожай будет хорошим, цены могут упасть и вы можете потерять все. С другой стороны, в случае серьезной засухи цены возрастут и будет получена существенная прибыль.

Случай третий. Управляющий крупного химического предприятия имеет много обязанностей. Ему необходимо удерживать на достаточно низком уровне затраты, обеспечивая вместе с тем производство большого количества продукции. Поскольку некоторые химикаты ядовиты, на предприятии введена в действие специальная система безопасности. Однако, несмотря на все прилагаемые усилия, управляющий все еще испытывает определенные сомнения в отношении того, насколько правдоподобной представляется крупная авария с опасными последствиями в будущем году. Такая ситуация оказывается очень неопределенной, и в средствах массовой информации время от времени появляются сообщения о подобных авариях. Для правильного понимания необходимых для поддержания безопасности затрат и оценки возможного выигрыша от их вложения принимается решение об исследовании вероятности аварии на предприятии.

Случай четвертый. Думали ли вы когда-нибудь о возможности выиграть в тотализатор? Это, наверное, представлялось вам очень маловероятным. Однако насколько это действительно невероятно? Ответ можно найти в газетной статье, в которой утверждается следующее.

"Вероятность выигрыша большого приза в тотализаторе, организованном по переписке журналом *Publishers Clearing House*, составляет 1 к 427 600 000. Это примерно в 300 раз более невероятно, чем попасть под удар молнии, вероятность чего составляет примерно 1 к 1 500 000 млн."[1]

Как же разобраться в неопределенных, содержащих элемент случайности, ситуациях таким образом, чтобы избежать влияния связанных с ними неточностей? Лучше всего начать с рассмотрения совершенно определенных, точных утверждений. Это может помочь узнать "достоверно", что же будет происходить (насколько это вообще можно узнать) даже в том случае, если в данной ситуации ни в чем нельзя быть полностью уверенным.

Такой подход устанавливает четкие границы. Рассмотрение начинается с четкого описания представляющего интерес процесса (*случайного эксперимента*). При этом составляется перечень всех возможных результатов (*выборочное пространство*). Может быть также ряд особых случаев, которые следует рассмотреть и которые либо будут наблюдаться, либо нет (например, "успешное выполнение проекта"); они называются *событиями*. Правдоподобие наступления события характеризуется конкретным числом, которое называют *вероятностью*.

При этом может возникнуть желание объединить информацию о более чем одном событии; например, оценить, насколько вероятно то, что откажут в работе *и* реактор, *и* система обеспечения безопасности. Может также возникнуть необходимость с течением времени обновлять имеющиеся значения вероятности с помощью так называемых *условных вероятностей*, отражающих доступную информацию.

[1] William P. Barrett "Bank on Lightning, Not Mali Contests", The Seattle Times, 1986, January 14, p. D1.

Самый простой способ решения таких вероятностных задач состоит в том, чтобы прежде всего построить *дерево вероятностей* для структурирования своих знаний о ситуации; это оказывается значительно проще, чем просто пытаться правильно комбинировать формулы. В некоторых случаях полезным оказывается также рассмотрение *таблиц совместных вероятностей* и *диаграмм Венна*, позволяющее углубить интуитивные представления и решить поставленную задачу.

6.1. Пример: за какой из дверей спрятан приз?

Представьте себе, что вы участвуете в игровом телевизионном шоу. Вас ждет приз, о котором вы мечтаете (путешествие на Гавайи или, может быть, отличная оценка на коллоквиуме?). Эта награда спрятана за одной из дверей — дверью №1, дверью №2 или дверью №3. За другими двумя дверьми нет ничего интересного. Вы видите три закрытые двери; какие-либо подсказки отсутствуют. Толпа участников передачи кричит, вдохновляя вас на подвиг; вы оцениваете ситуацию, собираетесь с мыслями и сообщаете присутствующим о своем выборе. Однако прежде чем открыть выбранную вами дверь, вам говорят, что сначала откроют другую, за которой приза нет. После этого вам дается возможность изменить свой выбор. Теперь закрытыми остались две двери: та, которую вы выбрали, и еще одна. Будете ли вы менять свой выбор? Толпа неистовствует, раздаются выкрики, призывающие вас изменить свое решение или не менять его. Оставите ли вы свой первоначальный выбор без изменения или выберете другую дверь? Подумайте над этим, а потом уже посмотрите ответ, который приведен и обоснован ниже[2].

* * * * *

Если вы решили сохранить свой первоначальный выбор, вы попадете в хорошую компанию — почти все студенты дают такой ответ. Однако, к сожалению, такой выбор будет ошибочным. Ваша способность принимать решения после знакомства с вероятностью несомненно улучшится.

Если вы решили изменить выбор — поздравляем[3]. Вы удвоили свой шанс выиграть с 1/3 до 2/3.

Принцип здесь состоит в том, что те, кто решился на изменение выбора, эффективно использовали новую полученную информацию. Те же, кто оставил свой выбор без изменений, не изменили и свою возможность выиграть. В чем состоит новая информация? Для того чтобы открыть дверь, которую вы не выбрали и за которой нет приза, устроители шоу должны, по меньшей мере, что-то знать о том, за какой дверью в действительности находится приз. Когда они открывают дверь, часть этой информации становится доступной и вам.

[2] Аналогичная задача описана в статье B. Nalebuff, "Puzzles: Choose a Curtain, Duel-ity, Two Point Conversions, and More", Journal of Economic Perspectives 1 (1987), p. 157–163. Позже эта задача была опубликована в колонке Marilyn Vos Savant журнала Parade, являющегося приложением ко многим воскресным газетам, и привлекла к себе внимание.

[3] А еще напишите, пожалуйста, для меня краткое сообщение о том, почему вы решили изменить свой выбор. Для меня представляют интерес различные интуитивные пути, которыми люди приходят к правильным ответам на поставленные перед ними вопросы. Пишите, пожалуйста: Andy Siegel, Department of Management Science, Box 353200, University of Washington, Seattle, Washington 98195. Спасибо!

Неформально пояснить ситуацию можно следующим образом. Представьте себе, что у вас есть двойник, который меняет выбор в то время, когда вы свой выбор не меняете. Поскольку осталось только две двери и приз должен находиться за одной из них, ваш двойник выиграет в каждом из случаев, когда вы проиграете. Поскольку ваш общий шанс на выигрыш остается неизменным и составляет 1/3, получается, что изменивший выбор двойник получает остальной шанс, равный 2/3. Для тех, кого это не убедило, в разделе 6.5 будет приведено детальное строгое решение.

Не стоит чрезмерно расстраиваться, если вам не удалось сделать оптимальный выбор. Подумайте лучше о том, насколько усилится ваша позиция в результате изучения вероятностей. Скорее даже следует испытывать оптимизм, поскольку в результате принимаемые вами решения окажутся лучше, чем у многих других людей, продолжающих "противиться переменам" даже после получения новой важной информации.

6.2. Как исследовать неопределенность

Построение строгой схемы для исследования случайных ситуаций мы начнем с тщательного определения и ограничения ситуации. Результатом такого определения будет *случайный эксперимент*, каждое выполнение которого приводит к некоторому *результату*, одному из списка возможных (так называемое *выборочное пространство*). Также мы будем иметь дело с рядом *событий*, каждое из которых может либо происходить, либо не происходить, в зависимости от результата случайного эксперимента.

Случайный эксперимент: *точное* определение случайной ситуации

Случайный эксперимент — это любая четко определенная процедура, дающая наблюдаемый результат, который нельзя точно предсказать заранее. Случайный эксперимент должен быть четко определен во избежание любых неожиданностей. Должна быть возможность однозначно фиксировать результат выполнения эксперимента. И наконец, результат такой процедуры не должен быть полностью предсказуемым заранее, поскольку в противном случае ситуация не является действительно случайной[4]. Ниже приведено несколько примеров случайного эксперимента, рассмотрение которых мы продолжим в этом разделе.

1. Планируется исследование дохода семей, проживающих вблизи места, где предполагается открыть новый ресторан. Случайным образом выбирается телефонный номер, по которому звонят в семью и записывают ее доход с точностью до доллара. Для того чтобы быть точным, необходимо четко определить результат в том случае, если никто не ответил на звонок или если

[4] Иногда результат может казаться полностью предсказуемым, как, например, в случае вопроса о том, будут ли занятия завтра или будет ли компания Ford по-прежнему лидировать в автомобильной промышленности в 2010 году. Однако и эти случаи также можно спокойно относить к числу "случайных экспериментов", поскольку некоторое сомнение в результате остается даже тогда, когда он представляется *почти* определенным.

собеседник не захотел ответить на заданный вопрос о доходе. Предположим, что в этом случае выбирается новый номер и звонки повторяются до тех пор, пока не будет получен конкретный ответ о величине дохода. В таком случае каждое выполнение случайного эксперимента будет давать в качестве результата конкретное число (размер дохода семьи). Несмотря на то что такой подход позволяет собрать данные для решения нашей вероятностной задачи, отсутствие ответов остается проблемой для статистического анализа, поскольку все еще существует некоторая группа (не ответивших на поставленный вопрос), для которой информация о семейном доходе отсутствует.

2. Отдел маркетинга планирует выбрать 10 типичных покупателей для формирования фокус-группы, в которой будет обсуждаться и выбираться один из семи предложенных вариантов дизайна новой лампы. (Результат — выбранный вариант дизайна.)
3. В соответствии с четко определенным процессом случайного выбора из завтрашней продукции выбирают пять замороженных обедов, готовят их и описывают их качество целым числом по шкале от 1 до 4. (Тщательность организации случайного процесса выбора гарантирует хороший эксперимент. Результатом является список оценок качества. Результат с полной определенностью неизвестен, поскольку в процессе любого производства возможны проблемы.)

Сложные ситуации обычно допускают много различных вариантов случайного эксперимента. При этом можно выбрать один или несколько наиболее подходящих вариантов. Так, например, выбор одного обеда для контроля качества сам по себе является случайным экспериментом, дающим в качестве результата одну оценку качества. Выбор для анализа пяти обедов также представляет собой случайный эксперимент (но большего масштаба), который в качестве результата дает уже список из пяти оценок качества.

Рассматривайте случайный эксперимент как сцену, на которой происходят различные события. Рассмотрение наблюдаемых событий в сравнительно меньших ситуациях позволяет более ясно увидеть возможные результаты.

Выборочное пространство: перечень возможных событий

Каждый случайный эксперимент характеризуется **выборочным пространством**, представляющим собой перечень *всех возможных результатов* этого эксперимента, описанным заранее, без знания о том, что действительно произойдет в ходе эксперимента. Обратите внимание на то, что в отношении выборочного пространства никакой неопределенности нет. Это вполне определенный список того, что может наблюдаться. Такой подход позволяет сделать случайную ситуацию более определенной и часто помогает также прояснить свое представление о ней. Ниже приведены выборочные пространства для описанных ранее случайных экспериментов.

1. В случае исследования дохода семей выборочное пространство представляет собой список возможных значений дохода. Предположим, что доход может принимать нулевое значение или быть положительным числом, и представим себе выборочное пространство как список неотрицательных значений денежных сумм в долларовом выражении.

$0
$1
$2
.
.
.
$34 999
$35 000
$35 001
.
.
.

2. Для фокус-группы, выбирающей дизайн новой лампы, выборочное пространство значительно меньше, оно состоит из семи предложенных вариантов дизайна:

Дизайн A	Дизайн E
Дизайн B	Дизайн F
Дизайн C	Дизайн G
Дизайн D	

3. В случае проверки качества замороженных обедов выборочное пространство представляет собой набор всех возможных вариантов оценок качества. Это набор списков, каждый из которых содержит 5 чисел (по одному на каждый из проверенных обедов) из интервала от 1 до 4 (каждое число характеризует качество соответствующего обеда). В данном случае выборочное пространство слишком велико для того, чтобы приводить его здесь полностью. Отметим, однако, что перечень элементов выборочного пространства может начинаться с варианта, содержащего все единицы (т.е. имеющего вид "1, 1, 1, 1, 1", что свидетельствует об отвратительном качестве всех проверенных обедов) и оканчиваться вариантом, содержащим все четверки (т.е. "4, 4, 4, 4, 4", что соответствует отменному качеству всей попавшей на проверку продукции). Между этими вариантами должны быть все возможные списки оценок качества обедов (например, "3, 2, 3, 3, 4")[5].

1	1	1	1	1	(первый список)
1	1	1	1	2	
1	1	1	1	3	
1	1	1	1	4	

[5] В данном случае выборочное пространство содержит 1024 варианта списков оценок. Это значение можно вычислить как $4\times4\times4\times4\times4=4^5$, поскольку существует 4 возможные оценки качества для первого обеда, 4 – для второго и так далее до 5-го выбранного для проверки обеда.

Окончание таблицы

1	1	1	2	1	
1	1	1	2	2	
1	1	1	2	3	
1	1	1	2	4	
1	1	1	3	1	
1	1	1	3	2	
1	1	1	3	3	
1	1	1	3	4	
1	1	1	4	1	
1	1	1	4	2	
1	1	1	4	3	
1	1	1	4	4	
1	1	2	1	1	
1	1	2	1	2	
1	1	2	1	3	
1	1	2	1	4	
1	1	2	2	1	
1	1	2	2	2	
1	1	2	2	3	
1	1	2	2	4	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
4	4	4	3	1	
4	4	4	3	2	
4	4	4	3	3	
4	4	4	3	4	
4	4	4	4	1	
4	4	4	4	2	
4	4	4	4	3	
4	4	4	4	4	(последний список)

Результат: что происходит в действительности

Каждый раз выполнение случайного эксперимента дает ровно один **результат.** Поскольку выборочное пространство содержит все возможные результаты, не должно быть никаких неожиданностей: результат эксперимента должен содержаться в выборочном пространстве.

Вот результаты одного выполнения каждого из рассматриваемых нами случайных экспериментов.

1. В исследовании дохода семьи после одного звонка, который остался без ответа (никого не оказалось дома), и одного, в результате которого был получен ответ "это не ваше дело," удалось дозвониться до человека, который назвал доход своей семьи (получен результат):

   $36 500

2. Для обсуждения в фокус-группе после длительных дискуссий о достоинствах различных вариантов дизайна лампы был сделан выбор (получен результат):

   Дизайн D

3. При исследовании качества замороженных обедов был получен список оценок качества для каждого из пяти проверенных обедов:

   3 3 1 2 2

События: они либо происходят, либо нет

В качестве формального определения **события** выступает любой набор указанных заранее, до проведения случайного эксперимента, результатов. Таким образом, можно сказать, что каждый раз при проведении случайного эксперимента мы наблюдаем, наступает определенное событие или нет. Можно определить для случайного эксперимента несколько событий или только одно. Каждое событие соответствует некоторому представляющему интерес свойству.

Ниже описаны некоторые события для каждого из рассматриваемых случайных экспериментов.

1. При исследовании дохода семьи можно рассмотреть три разных события.

Первое событие	Низкий уровень дохода*	От $10 000 до $24 999
Второе событие	Средний уровень дохода	От $25 000 до $44 999
Третье событие	Достаточный уровень дохода	$20 000 и выше

* Список результатов для события "Низкий уровень дохода" может быть расписан более детально как $10000, $10001, $10002, ..., $24997, $24998, $24999.

   Для некоторого наблюдаемого результата $36 500 мы видим, что первое событие не наступает, в то время, как два других наступают. Таким образом, эта семья имеет "средний уровень дохода" и "достаточный уровень дохода", но не "низкий уровень дохода".

2. Для случая работы фокус-группы по обсуждению дизайна лампы рассмотрим событие "выбранный дизайн удобен для производства", включающее следующие варианты дизайна:

Дизайн A
Дизайн B
Дизайн C
Дизайн F

Поскольку наблюдаемый результат, дизайн D, в число этих вариантов не входит, данное событие "не произошло". К сожалению, фокус-группа выбрала дизайн, который достаточно сложен для производства.

3. Для проверки качества замороженных обедов рассмотрим событие "все образцы — хорошего или высшего качества". Это означает, что все пять обедов должны быть оценены значениями 3 или 4. При этом, например, наблюдаемый результат

   3 3 1 2 2

   не может быть квалифицирован как "все образцы — хорошего или высшего качества" (три последних обеда получили слишком низкие оценки). Таким образом, данное событие *не произошло*. Однако если бы результат имел вид

   3 4 3 4 4,

   то рассматриваемое событие имело бы место (поскольку все обеды имели бы качество 3 или 4).

   Чтобы дать этому событию формальное определение, т.е. определить его в качестве набора тех результатов, при получении которых событие имеет место, необходимо составить следующую таблицу:

3	3	3	3	3	(первый список)	4	3	3	3	3	
3	3	3	3	4		4	3	3	3	4	
3	3	3	4	3		4	3	3	4	3	
3	3	3	4	4		4	3	3	4	4	
3	3	4	3	3		4	3	4	3	3	
3	3	4	3	4		4	3	4	3	4	
3	3	4	4	3		4	3	4	4	3	
3	3	4	4	4		4	3	4	4	4	
3	4	3	3	3		4	4	3	3	3	
3	4	3	3	4		4	4	3	3	4	
3	4	3	4	3		4	4	3	4	3	
3	4	3	4	4		4	4	3	4	4	
3	4	4	3	3		4	4	4	3	3	
3	4	4	3	4		4	4	4	3	4	
3	4	4	4	3		4	4	4	4	3	
3	4	4	4	4		4	4	4	4	4	(последний список)

   Таким образом, событие "все образцы — хорошего или высшего качества" наблюдается только в том случае, если полученный результат входит в данный полный список. Посмотрите, есть ли в нем результат "3, 4, 3, 4, 4"?

Как видите, здесь нет ничего загадочного. Необходимо тщательно определить случайный эксперимент, описать выборочное пространство, результаты, а также

события — и неопределенную ситуацию удается привести в надлежащий порядок. И наконец, события позволяют формулировать обычные содержательные утверждения, как, например, "Сегодняшний товар прошел контроль качества" или "Эй, Марджи, я нашел еще одну семью с достаточным уровнем дохода!"

6.3. Насколько вероятно событие?

Как нам уже известно, каждый раз при проведении случайного эксперимента каждое из событий либо происходит, либо не происходит. Это пока еще мало о чем говорит. Нам хотелось бы знать, насколько *правдоподобно* конкретное событие. Эта правдоподобность характеризуется числом, которое называется *вероятностью события*. Сейчас мы дадим определение этого понятия, обозначим, откуда берутся числа, характеризующие вероятность, а также покажем, как вероятность указывает приблизительное количество появлений некоторого события при многократном повторении случайного эксперимента.

Каждое событие имеет свою вероятность

Каждому событию соответствует число в интервале от 0 до 1, которое называют **вероятностью** события. Вероятность показывает, насколько правдоподобно наступление данного события при каждом выполнении случайного эксперимента. Вероятность, равная 0, означает, что соответствующее событие по сути никогда не наблюдается, а вероятность, составляющая 1, свидетельствует о том, что данное событие происходит практически всегда[6]. В целом вероятность показывает приблизительную долю случаев, в которых ожидается наступление события.

Вероятность	Интерпретация
1,00	Событие наблюдается (практически) всегда
0,95	Событие наблюдается примерно в 95% случаев (оно очень вероятно)
0,50	Событие наблюдается примерно в половине случаев
0,37	Событие наблюдается примерно в 37% случаев
0,02	Событие наблюдается примерно в 2% случаев (оно маловероятно, но возможно)
0,00	Событие не наблюдается (практически) никогда

Интерпретация этих чисел требует определенной осторожности. Если вы можете провести только один эксперимент и повторить его нельзя (например, если речь идет о вероятности того, что старое здание будет успешно взорвано), вероятность в 0,97 свидетельствует о большой вероятности успеха с малой возможностью потерпеть неудачу. Однако неверно будет в таком случае утверждать, что "успех ожидается в 97% случаев", если только вы не собираетесь повторить эту процедуру в дальнейшем для многих подобных старых строений.

[6] Технические детали математического определения таковы, что событие, имеющее вероятность 0, все же может в действительности произойти. Однако такое событие может наблюдаться настолько редко, что принято говорить о том, что оно "практически никогда не происходит". Если вас интересует этот вопрос, попробуйте, например, представить себе вероятность того, что завтра будет добыто именно 254896,3542082947839478... баррелей нефти.

Можно также определить вероятность, используя такое понятие, как *шанс*. Шанс представляет собой положительное число, равное частному от деления вероятности наступления события на вероятность того, что это событие не произойдет:

Шанс

$$\text{Шанс} = \frac{\text{Вероятность наступления события}}{\text{Вероятность того, что событие не произойдет}} = \frac{\text{Вероятность}}{1 - \text{Вероятность}}.$$

Так, например, вероятность 0,5 соответствует шансу 0,5/(1−0,5)=1. Это иногда формулируется в виде "шансы 1 к 1". Вероятность 0,8 соответствует шансу 0,8/(1−0,8)=4, или 4 к 1. Большой шанс соответствует более высокой вероятности и большему правдоподобию. Обратите внимание, что, несмотря на то, что вероятность не может выходить за пределы промежутка от 0 до 1, шанс может принимать любое неотрицательное значение.

Откуда берутся значения вероятности

В учебниках числа, являющиеся значениями вероятности, часто приводятся в условии задачи наряду с другой информацией. В реальной жизни пометки вида "Кстати, вероятность того, что болт окажется дефектным и приведет к аварии, равна 0,003", естественно, отсутствуют. Где же взять характеризующие вероятность числа для использования в реальной жизни? Для этого есть три основных способа: найти *относительную частоту* (с помощью эксперимента), вычислить *теоретическое значение вероятности* (используя формулы) или воспользоваться *субъективной оценкой вероятности* (на основе суждений).

Относительная частота и закон больших чисел

Предположим, что у нас есть возможность провести случайный эксперимент неограниченное число раз в совершенно одинаковых, за исключением влияния случайного фактора, условиях. **Относительная частота** некоторого события равна отношению количества раз наступления этого события ("числа появлений события") к общему количеству проведенных экспериментов ("числу повторов"). Эта величина может задаваться либо в виде дроби (например, 0,148), либо в процентном выражении (например, 14,8%). Формула для вычисления относительной частоты события имеет следующий вид:

$$\text{Относительная частота события} = \frac{\text{Число появлений события}}{\text{Число повторов случайного эксперимента}}.$$

Например, если вы опросили 536 человек, которые согласились ответить на ваш вопрос, и нашли, что в 212 случаях семьи имеют достаточный доход размером $20000 или более, относительная частота будет равна:

$$\frac{212}{536} = 0,396, \text{ или } 39,6\%.$$

Величина относительной частоты 39,6% определена для случайного эксперимента "выяснение дохода у человека, согласившегося ответить на вопрос" с представляющим интерес событием "доход составляет $20000 или более". Если использовать такие определения, становится понятным, что в данном случае проведен случайный эксперимент, причем он повторен n = 536 раз. В данном случае можно также вести речь и о другом случайном эксперименте, который также может представлять интерес, а именно: "определение дохода семей 536 человек, согласившихся ответить на вопрос". Однако для этого более масштабного случайного эксперимента относительная частота не имеет смысла, поскольку этот эксперимент проведен только один раз.

Эти различия не тривиальны. Менеджеры часто считают, что подобная "строгость рассуждений" полезна лишь в случае более сложных проблем. Относительная частота и вероятность некоторого события — это близкие, но разные понятия. Существенное различие состоит в том, что вероятность представляет собой *точное число*, в то время как относительная частота — это *случайное число*. Это связано с тем, что вероятность события оказывается свойством базовой ситуации (случайного эксперимента, выборочного пространства, события), а относительная частота зависит от (случайных) результатов, получаемых при проведении случайного эксперимента n раз.

Относительные частоты могут использоваться для оценки (наилучшая приблизительная оценка) значения вероятности в случае наличия информации, основанной на предшествующем опыте. Так, например, величину относительной частоты (в данном случае величину 39,6% для достаточного уровня дохода) можно использовать в качестве приближения для истинного значения вероятности того, что при случайном выборе у человека, согласившегося ответить на вопрос, окажется достаточный уровень дохода. Таким образом, данное значение, 0,396, можно использовать так, как будто это и есть вероятность. Не следует, однако, забывать о том, что существует различие между истинной (неизвестной) величиной вероятности и наилучшей приблизительной оценкой, полученной на основе рассмотрения относительной частоты.

Закон больших чисел гласит, что относительная частота должна быть близкой к вероятности, если эксперимент проведен много раз, т.е. при больших значениях n. Насколько, как правило, будет близка (случайная) величина относительной частоты к (фиксированной) величине вероятности? Ответ, который зависит от того, насколько правдоподобно данное событие и сколько раз повторен эксперимент (n), приведен в табл. 6.3.1 с использованием понятия стандартного отклонения. Если, например, вероятность события составляет 0,75 и случайный эксперимент повторяется n = 100 раз, можно ожидать, что относительная частота окажется в среднем приблизительно на 0,04 выше или ниже истинной вероятности (0,75).

На рис. 6.3.1 и 6.3.2 показан пример того, как относительная частота (подверженная случайным колебаниям ломаная линия на каждом из графиков) выступает в качестве приближения для вероятности (прямая горизонтальная линия на высоте 0,25 на каждом графике). Обратите внимание, что на рис. 6.3.2 масштаб по вертикали растянут. Это сделано потому, что при больших n относительная вероятность ближе к действительному значению вероятности.

Таблица 6.3.1. Стандартное отклонение относительной частоты

	Значение вероятности 0,50	Значение вероятности 0,25 или 0,75	Значение вероятности 0,10 или 0,90
n = 10	0,16	0,14	0,09
25	0,10	0,09	0,06
50	0,07	0,06	0,04
100	0,05	0,04	0,03
1000	0,02	0,01	0,01



Рис. 6.3.1. Относительная частота события выступает в качестве аппроксимации для вероятности этого события (в данном случае 0,25) и в целом приближается к ней с ростом n. При n = 100 ожидается разница не более 0,04



Рис. 6.3.2. Относительная частота события, показанная для числа повторов случайного эксперимента n от 100 до 1000. В большинстве случаев при n = 1000 отклонение относительной частоты от вероятности (в данном случае от 0,25) не превышает 0,01. Обратите внимание на изменение масштаба по сравнению с предыдущим графиком

Поскольку относительная частота достаточно близка, по крайней мере для больших значений n, к вероятности (в соответствии с законом больших чисел), относительную частоту появления некоторого события можно использовать в качестве "наилучшей приблизительной оценки" (на основе имеющихся данных) вероятности этого события.

Пример. *Насколько велика изменчивость выпущенной сегодня продукции высшего качества*

Предположим, что вы проверили качество 50 единиц из сегодняшнего выпуска продукции вашей фирмы. Из предыдущего опыта вам известно, что в течение длительного периода времени 25% произведенной продукции оценивались как продукция высшего качества и считались годными для продажи на экспорт самым требовательным иностранным заказчикам. Чего следует ожидать на сегодняшний день? Получится ли ровно 25% из этих 50, т.е. 12,5 единиц продукции высшего качества? Естественно, нет; однако насколько ниже или выше этого значения можно ожидать реальный результат?

Стандартное отклонение относительной частоты (0,06, или 6%) из табл. 6.3.1 поможет получить ответ на этот вопрос. Давайте рассуждать таким образом. Сегодня 50 раз выполнен случайный эксперимент "производство единицы продукции и определение ее качества"[7]. Относительная частота события "единица продукции имеет высшее качество" будет равна проценту продукции высшего качества в сегодняшнем производстве. Эта относительная частота будет близка к вероятности (25%, или 0,25) и будет (согласно данным таблицы) отстоять от значения 0,25 приблизительно на 0,06, или 6%. Если перейти от процентов к единицам продукции, эта величина изменчивости составит $0,06 \times 50 = 3$ единицы товара. Таким образом, вы ожидаете получить результат "12,5 плюс или минус 3" единиц продукции высшего качества.

Изменчивость, которая характеризуется 3 единицами продукции, интерпретируется так, как это обычно делается для стандартного отклонения. Не следует удивляться, если среди всей продукции окажется 10 или 14 единиц высшего качества или даже 7 или 18 (примерно две величины стандартного отклонения от 12,5). Однако окажется странным и неприятным, если таких товаров окажется всего лишь несколько штук. Подобный результат может свидетельствовать о необходимости вносить коррективы в процесс производства. Будет также удивительно, хоть и очень приятно, обнаружить 23 единицы продукции высшего качества (пора открывать шампанское и изо всех сил стараться помочь рабочим вспомнить, как это им удалось!).

Теоретическое значение вероятности

Теоретическое значение вероятности — это значение вероятности, которое вычисляют с использованием точной формулы, основанной на математической теории или модели. Этот подход можно использовать только для систем, имеющих строгое математическое описание. Большинство методов вычисления теоретического значения вероятности слишком сложны, и мы не будем их касаться. Рассмотрим только особый случай применения правила *равной вероятности*, а в главе 7 — специальные распределения вероятности, такие, как нормальное и биномиальное распределения.

Правило равной вероятности

Если все результаты оказываются одинаково вероятны — а в этом обычно существуют определенные сомнения, — легко найти вероятность любого события, воспользовавшись теоретическим подходом. Вероятность пропорциональна

[7] В данном исследовании предполагается, что единицы продукции производятся независимо друг от друга.

числу результатов, которые образуют данное событие, и вычисляется по следующей формуле.

Если все результаты равновероятны, то

$$\text{Вероятность события} = \frac{\text{Число результатов в событии}}{\text{Общее количество возможных результатов}}$$

Пример. *Подбрасывание монеты и тасование карт*

Поскольку подброшенная монета, скорее всего, упадет либо орлом, либо решкой вверх, вероятность каждого из этих событий составляет 1/2. А что можно сказать о трех результатах, "орел," "решка" и "станет на ребро"? Поскольку в данном случае рассматриваются три события, означает ли это, что вероятность каждого из событий равна 1/3? Естественно, нет; ведь в этом случае явно нарушается требование "все результаты должны быть равновероятны". Правило равной вероятности оказывается неприменимым, поскольку вероятность того, что подброшенная монета при падении станет на ребро, очень мала по сравнению с вероятностями двух других возможностей.

Поскольку карты перед игрой обычно тасуют, что помогает обеспечить случайное расположение карт в колоде, правило *равной вероятности* должно быть применимо к любой попытке указать свойство произвольной карты. Так, например, поскольку 13 из 52 карт колоды — это черви, вероятность того, что одна выбранная случайным образом карта будет червовой масти составляет 0,25. Аналогичным образом находим, что вероятность получения туза равна 4/52 = 7,7%, а вероятность того, что случайная карта окажется джокером, составляет 2/52 = 3,8%.

Пример. *Возможности трудоустройства для мужчин и женщин*

Предположим теперь, что 15 имеющих одинаковую квалификацию человек подали заявления на некоторую вакансию, причем 6 претендентов — женщины. Если в этой группе выбор при приеме на работу осуществляется случайным образом (и, таким образом, пол претендента не учитывается), вероятность того, что предпочтение будет отдано женщине, составляет 6/15 = 0,40, или 40%. Эта ситуация соответствует правилу равной вероятности, срабатывающему в случае проведения случайного эксперимента по выбору одного из претендентов, с событием "выбор пал на женщину". Данное событие включает 6 результатов (6 женщин), а вероятность составляет 6 из 15, полного числа претендентов на должность (общего количества результатов).

Пример. *Получение исходных материалов*

Представьте себе, что ваш поставщик имеет склад, в котором находится 83 коробки передач, из них 2 имеют дефекты. Если одна из них выбирается случайным образом (какое из изделий имеет дефекты, неизвестно), чему равна вероятность того, что вы получите дефектное изделие? Ответ в соответствии с правилом равной вероятности составляет 2 из 83, или 2,4%.

Субъективная оценка вероятности

Субъективная оценка вероятности получается на основе суждения определенного лица о вероятности некоторого события. Такой подход может показаться не очень научным, однако часто оказывается, что это и есть лучшее, что можно сделать в отсутствие предыдущего опыта (т.е. не имея возможности использовать относительную частоту) и в отсутствие соответствующей теории (т.е. без возможности вычислить теоретическое значение вероятности). Один из путей улучшения качества подхода на основе субъективной оценки вероятности состоит в использовании мнения эксперта в данной области. Например, можно воспользоваться мнением специалиста по банковским инвестициям для оценки вероятности того, что слияние конкурирующих фирм окажется успешным, или мнением инженера о технической осуществимости нового технологического подхода в области энергетики.

Пример. *Ведение судебного процесса*

Против вашей фирмы возбуждено дело в суде. Ваша первая реакция — "О, только не это!" Однако впоследствии вы осознаете, что на фирмы подают в суд по разным вопросам очень часто. Так что вы успокаиваетесь и беретесь за ознакомление с деталями иска, в число которых входит требование возмещения опосредованного ущерба в сумме $5 000 000. Вы созываете собрание административных работников своей фирмы и юрисконсультов для обсуждения стратегии. Выбор наиболее эффективной стратегии оказывается связанным с оценкой того, насколько правдоподобными оказываются различные возможные исходы дела.

Это, естественно, ситуация, в которой присутствуют вероятности различных событий. Вы оказались в самом центре масштабного случайного эксперимента.

Давайте проследим возможное развитие судебного процесса и перечислим возможные результаты в виде (1) потраченных сумм (на оплату издержек и возмещение ущерба, если таковые есть) и (2) разрешения процесса (его прекращение, разрешение без судебного разбирательства, вынесение решения судьей или судом присяжных).

Поскольку многие судебные разбирательства оканчиваются до вынесения решения по делу, вначале рассмотрим различные возможные затраты на разрешение процесса мирным путем. Возможны следующие представляющие интерес события.

1. Урегулирование вопроса без больших затрат: менее $100 000.
2. Урегулирование вопроса со средними затратами: от $100 000 до $1 000 000.
3. Урегулирование вопроса с большими затратами: более $1 000 000.

Как определить вероятность этих событий? Достаточное количество аналогичных случаев, которое позволило бы применить подход на основе относительных частот событий, отсутствует — даже после обращения к компьютерным базам данных с целью нахождения примеров подобных дел. Научной теории, которая позволила бы воспользоваться формулой для расчета таких вероятностей, также нет, что лишает возможности применить метод теоретического вычисления вероятности. Остается единственный путь — провести субъективную оценку вероятности.

Для субъективной оценки вероятности перечисленных событий собрание принимает решение обратиться ко мнению юрисконсульта, хорошо знающего судебные процессы такого типа. Изучив конкретные детали дела и оценив его на основе сравнения с несколькими имевшими место ранее аналогичными случаями, а также учитывая сегодняшние тенденции судебной системы, юрисконсульт представил следующие значения субъективной вероятности.

1. Урегулирование вопроса без больших затрат: 0,10.
2. Урегулирование вопроса со средними затратами: 0,65.
3. Урегулирование вопроса с большими затратами: 0,15.

Обратите внимание на то, что в сумме эти вероятности дают 0,90, при этом остается вероятность в 10%, что дело будет решаться судом (т.е. урегулирование без судебного рассмотрения не состоится).

Эти величины субъективной вероятности представляют наилучшую из доступных вам оценок правдоподобия различных вариантов развития событий, и из них следует, что, вероятно, вопрос можно будет урегулировать со средними затратами. Вам становится легче от того, что наиболее вероятная сумма, необходимая для урегулирования, существенно меньше $5 000 000, о которых шла речь вначале. Теперь можно воспользоваться этими значениями вероятности для того, чтобы принять непростые решения относительно типа и объема доводов, которые можно привести в свою защиту.

Анализ методом Байеса и частотный анализ

В применяемом в статистическом анализе **методе Байеса** субъективные вероятности используются формально, в математических вычислениях. На рис. 6.3.3 (вверху) показано, как с помощью метода Байеса наблюдаемые данные используются вместе с априорными вероятностями и с моделью (математическим описанием ситуации) для вычисления результатов.

Существует также другой подход, называемый **частотным анализом**. На первый взгляд такой подход представляется более объективным, поскольку соответствующие вычисления основываются только на наблюдаемых данных и модели. Однако при более внимательном рассмотрении (рис. 6.3.3, внизу) становится понятна замаскированная роль субъективных суждений, определяющих выбор плана исследования (который и приводит к появлению данных) и выбор модели.

Анализ на основе метода Байеса может оказаться полезным в сложных статистических исследованиях в бизнесе, особенно в тех случаях, когда в исследовании участвует эксперт, предоставляющий основанную на беспристрастных суждениях достоверную информацию, которая включается в исследование наряду с существующими данными. Если такую экспертную оценку можно эффективно использовать при планировании исследования и выборе модели, то необходимости в проведении сложных вычислений по методу Байеса не возникает. Однако в случае достаточно важного исследования и достаточно точной экспертной оценки следует рассмотреть вопрос и об использовании метода Байеса.

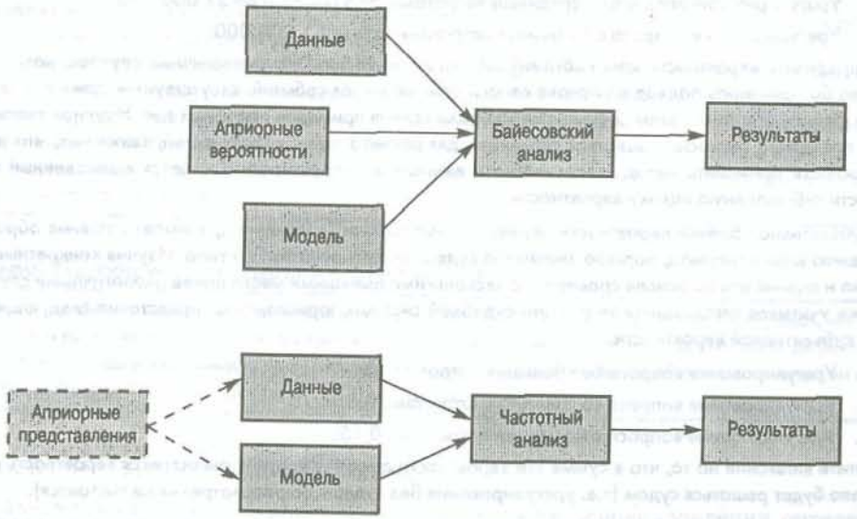


Рис. 6.3.3. Общие подходы Байесовского и частотного методов статистического анализа. В обоих подходах применяется априорная субъективная информация. Байесовский подход основан на непосредственном использовании этой информации для вычисления результатов. В частотном анализе априорная информация используется неформально, чтобы создать основу для более "объективных" вычислений

6.4. Как совместить информацию о нескольких событиях

Понимание вероятностей дает два больших преимущества. Во-первых, то, о чем мы уже говорили: оценка правдоподобия события как выраженной числом вероятности и определение соответствующего случайного эксперимента. Это большой шаг на пути к пониманию неясной, неопределенной ситуации. Второе преимущество состоит в возможности рассмотрения *комбинаций событий* и определении того, как, используя знания вероятностей некоторых событий, можно определить вероятности других событий, представляющих, возможно, больший интерес и имеющих большее значение.

Итак, наша цель заключается в том, чтобы из имеющейся у нас информации с помощью основных правил работы с вероятностями получить новую информацию. Если говорить строго, такая результирующая информация не будет действительно "новой", поскольку она логически вытекает из уже имеющейся. Умение быстро делать такие логические выводы поможет использовать сценарии вида *что будет, если...* для принятия бизнес-решений.

Диаграммы Венна позволяют увидеть все возможности

Диаграмма Венна — это рисунок, содержащий все множество возможных результатов (выборочное пространство) в виде прямоугольника с размещенными в нем событиями, часто в виде кругов или овалов, подобно тому, как это показано на рис. 6.4.1. Каждая точка внутри прямоугольника представляет возможный результат. Каждое выполнение случайного эксперимента приводит к случайному выбору одной из точек (получению результата); если эта точка попадает в обозначающий событие круг, то данное событие "произошло". В противном случае событие не происходит.

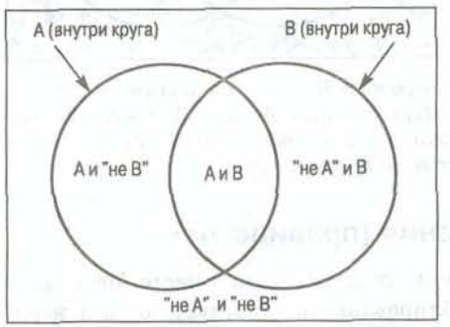


Рис. 6.4.1. Диаграмма Венна с двумя событиями, каждое из которых показано в виде круга. Обратите внимание на то, что часть точек (результатов) попадает в оба круга, часть находится вне обоих кругов, а часть располагается только в одном круге и не принадлежит другому. Это удобное наглядное представление всех возможностей для случая двух событий

Не событие

Дополнением некоторого события (также часто используют термин **противоположное** событие. — *Прим. ред.*) называют такое событие, которое происходит только в случае, когда первое событие *не* происходит. Каждое событие имеет свое дополнение.[8] Ниже приведено несколько примеров событий и их дополнений.

Событие	Дополнение события
Успешный выход товара на рынок	Неуспешный выход
Рост курса акций	Курс акций не меняется или падает
Товар имеет приемлемое качество	Качество товара неприемлемо

Если событие представляет собой некоторый набор результатов, то его дополнение включает все те результаты выборочного пространства, которые *не* входят в этот набор. Это иллюстрирует диаграмма Венна на рис. 6.4.2.

При построении дополнения события необходимо убедиться в том, что рассмотрены все возможные результаты. Например, дополнение к событию "рост цен" описывается *не* как "снижение цен", поскольку необходимо также учесть и возможность отсутствия изменения цен.

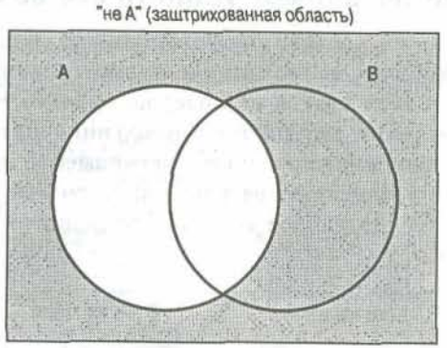


Рис. 6.4.2. Диаграмма Венна с показанным на ней дополнением "не А". Дополнение события А представляет собой событие, которое происходит тогда (и только тогда), когда событие А не происходит

Правило дополнения (правило *не*)

Поскольку событие и его дополнение вместе представляют все возможные результаты, причем дублирование отсутствует, сумма вероятностей этих двух событий равна 1. Таким образом, для события, которое мы обозначим А, и его дополнения, обозначенного "не А", мы получаем следующее правило:

вероятность А + вероятность "не А" = 1.

[8] Если вы встретите это слово на английском языке, необходимо обратить внимание на его написание. Каждое событие имеет "дополнение" (complement), но только те события, которые для нас приятны, можно "приветствовать" (compliment).

Эта формула позволяет определить вероятность дополнения события.

Правило дополнения

вероятность "не А" = 1 – вероятность А.

Если, например, известно, что вероятность успешного выпуска товара на рынок равна 0,4, вероятность дополнительного события "неуспешный выпуск товара на рынок" составляет 1 – 0,4 = 0,6.

Правило дополнения — первый пример использования метода получения "новой" информации о вероятности (в данном случае — о вероятности дополнительного к исходному события) из фактов, которые уже известны (вероятность исходного события). Если это вас еще не впечатлило — дальше вы увидите, что этот метод дает много новых результатов.

Одно событие *и* другое

Любые два события можно объединить и таким образом получить новое событие, которое называют их **пересечением** (часто также используют термин **произведение** событий. — *Прим. ред.*). Пересечение двух событий наблюдается каждый раз, когда оба события, *и* одно *и* другое, происходят в результате выполнения случайного эксперимента.

Когда два события рассматриваются в виде наборов результатов, их пересечение представляет собой новый набор, включающий все те результаты, которые входят одновременно в оба исходных набора. Пересечение показано на рис. 6.4.3.

Представим себе, например, что менеджер фирмы, осуществляющей коммерческие поставки некоторого товара, рассматривает вопрос о том, что следует делать, если в следующем году будет наблюдаться спад в экономике и конкуренты отреагируют на него снижением цен. Случайный эксперимент в этом случае описывается так: "в конце следующего года рассматриваем и регистрируем состояние экономики и ценовую политику конкурентов". Два события здесь — это "спад" (который либо будет наблюдаться, либо нет) и "снижение цен конкурен-

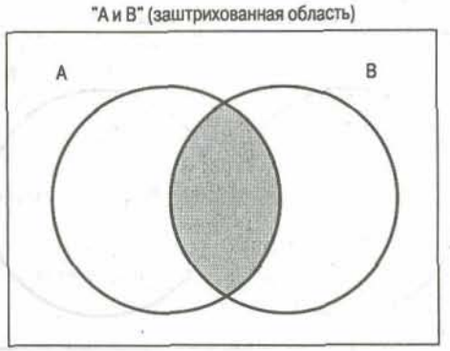


Рис. 6.4.3. Диаграмма Венна для пересечения событий "А и В". Это событие происходит тогда (и только тогда), когда в результате одного выполнения случайного эксперимента происходят и событие А, и событие В

тами". Нас интересует новое событие "спад *и* снижение цен конкурентами". Для разработки политики полезно рассмотреть, насколько правдоподобна такая новая возможность развития ситуации.

Если два события не могут наблюдаться вместе

Два события, которые не могут происходить одновременно, называются взаимоисключающими, или **несовместимыми между собой событиями**. Нельзя, например, завершить год с получением одновременно "очень высокой" и "очень низкой" прибыли. Невозможно также, чтобы ваш следующий принятый на работу сотрудник оказался одновременно "мужчиной, представителем национального меньшинства, с докторской степенью" и "женщиной, инженером и членом бейсбольной ассоциации". На рис. 6.4.4 показана диаграмма Венна для двух несовместимых событий. Она имеет вид двух непересекающихся окружностей.

Правило пересечения (*и*) для несовместимых событий

Поскольку два несовместимых события не могут происходить одновременно, вероятность их пересечения равна нулю. Проблемы с определением пересечения двух несовместимых событий не возникает: это пересечение вполне справедливо можно рассматривать как такое событие, которое никогда не происходит. Поэтому вероятность такого события равна нулю.

Правило пересечения событий (*и*) для двух несовместимых событий

Вероятность "A и B" = 0

Одно событие *или* другое

Любые два события можно объединить и получить новое событие, которое называют их **объединением** (часто также используют термин **сумма** событий. — *Прим. ред.*). Такое событие наблюдается каждый раз, когда одно *или* другое исходное событие (или оба эти события) происходят в результате одного выполнения случайного эксперимента[9].



Рис. 6.4.4. Диаграмма Венна для двух несовместимых событий, которые не могут произойти одновременно в результате одного случайного эксперимента. Поскольку круги не пересекаются, точек, общих для обоих событий, не существует

Если два события представлены наборами результатов, то объединение этих событий представляет собой новый набор, состоящий из всех результатов, содержащихся в каждом (или обоих) исходных наборах. Иллюстрация этого утверждения приведена на рис. 6.4.5.

Представим себе, например, что срок трудового контракта рабочих вашей фирмы подходит к концу и вы ожидаете, что они потребуют большой добавки к зарплате и дополнительных льгот. Доверяя предложенному менеджерами наиболее правдоподобному варианту нового контракта, вы хотите оценить его возможные последствия, чтобы предусмотреть возникновение непредвиденных обстоятельств. В частности, ваши рабочие могут объявить забастовку. Другая возможность состоит в спаде рабочей активности. Любое из таких событий создаст проблемы. Случайный эксперимент в данном случае состоит в том, чтобы "подождать и зарегистрировать реакцию рабочих на предложенный менеджерами проект контракта". Два рассматриваемых события здесь — это "забастовка" и "снижение темпа работы". Понятно, что объединение этих двух событий, "забастовка или снижение темпа работы", отражает тот широкий круг проблем, которые потребуют вашего внимания.

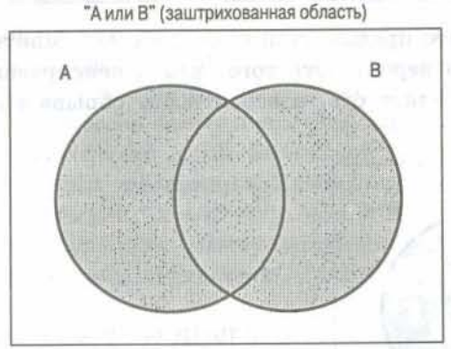


Рис. 6.4.5. Диаграмма Венна, иллюстрирующая объединение событий "A или B". Это событие наблюдается каждый раз, когда в результате одного выполнения случайного эксперимента происходит либо событие A, либо событие B, либо оба эти события

Правило объединения (*или*) для несовместимых событий

Если два события несовместимы (т.е. они не могут произойти одновременно), можно точно определить вероятность их объединения, сложив вероятности этих двух событий.

[9] Мы будем использовать союз "или" для обозначения "одно, или другое, или оба". Таким образом, можно сказать, что событие "инфляция или спад" наблюдается не только исключительно в периоды чистой инфляции или чистого спада, но и в течение тех сравнительно редких периодов, когда в экономике преобладают обе эти тенденции.

Поскольку два несовместимых события не имеют результатов, которые могли бы наблюдаться одновременно, при сложении вероятностей этих событий каждый из результатов учитывается только один раз. Таким образом, вероятность объединения двух несовместимых событий представляет собой сумму вероятностей каждого из них.

Правило объединения (*или*) для несовместимых событий

вероятность "A или B" = вероятность A + вероятность B.

Нахождение *или* из *и* и наоборот

Если известны вероятности трех различных событий, A, B и "A и B", можно найти вероятность "A или B". Эта вероятность находится сложением двух вероятностей базовых событий с последующим вычитанием вероятности их пересечения. Вычитание исключает те результаты, которые при сложении учитываются *два раза*, как показано на рис. 6.4.6. Вероятность события "A или B" выражается *следующим соотношением:*

Нахождение *или* из *и*

вероятность "A или B" = вероятность A + вероятность B – вероятность "A и B".

Представим себе, что предыдущий опыт работы ремонтной мастерской свидетельствует о том, что вероятность того, что в неисправном приборе перегорел предохранитель, составляет 6%, а вероятность обрыва провода составляет 4%.

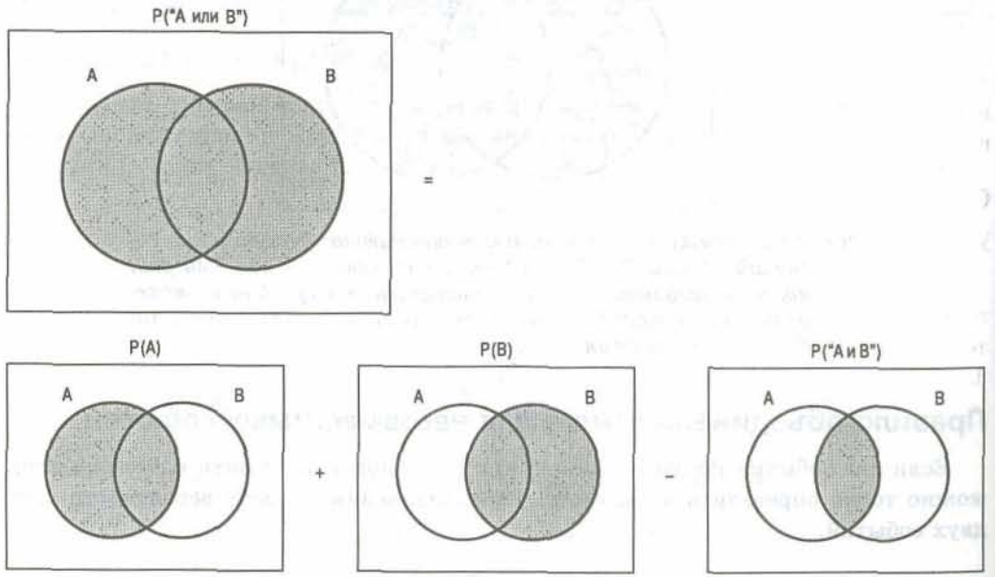


Рис. 6.4.6. Диаграмма Венна, иллюстрирующая вычисление вероятности события "A или B". Сначала складываем вероятность события A и вероятность события B. Затем вычитаем то, что при сложении было учтено дважды, а именно вероятность события "A и B"

Предположим также, что в 1% случаев всех обращений приборы поступали в ремонт с перегоревшим предохранителем *и* обрывом провода. Обладая этими сведениями, можно легко найти вероятность того, что в конкретном сданном в ремонт приборе присутствует одна из этих неисправностей (или обе).

Вероятность того, что "перегорел *предохранитель или есть обрыв провода*" равна 0,06 + 0,04 – 0,01 = 0,09.

Таким образом, в случае 9% обращений в мастерскую прибор имеет одну из этих неисправностей (или обе сразу).

Путем алгебраических преобразований можно также найти выражение для вычисления *вероятности* "A и B" *через вероятности* A, B и "A или B":

Нахождение *и* из *или*

вероятность "A и B" = вероятность A + вероятность B – вероятность "A или B".

Таким образом, зная любые три из этих четырех вероятностей (вероятностей событий A, B, "A и B" и "A или B"), можно найти четвертую, неизвестную вероятность.

В каких случаях оказываются полезными эти формулы? Один из случаев их применения состоит в том, чтобы взять в качестве исходной известные сведения о вероятностях и вычислить по соответствующей формуле вероятность другого события, возможно, представляющего больший интерес или имеющего большую важность. Другой случай — если мы хотим убедиться, что информация, на которой основываются решения, логически непротиворечива. Предположим, например, что у нас есть вероятности для событий A и B, вычисленных как относительные частоты на основе данных прошлых наблюдений. Планируется использовать субъективную оценку вероятности событий "A и B" и "A или B". При этом может оказаться полезным убедиться в том, что *связь между четырьмя рассматриваемыми* величинами вероятности не противоречит приведенным выше формулам.

Одно событие *при условии* другого: учет имеющейся информации

Рассматривая *задачу определения вероятности наступления события* с учетом того, что некоторое другое событие уже произошло, мы приходим к понятию **условной вероятности** первого события *при условии наступления* второго события. (Все простые вероятности, о которых шла речь до сих пор, можно назвать безусловными вероятностями — это позволит избежать излишней путаницы.) Приведем несколько примеров условных вероятностей.

1. Предположим, ваша местная команда может выиграть важную игру с вероятностью 70%. Теперь введем в рассмотрение новую информацию, соответствующую событию "после окончания первого тайма команда выигрывает". В зависимости от того, реализуется ли это событие, вероятность победы изменяется. Вероятность победы команды при условии, что она действительно выигрывает после первого тайма, окажется выше и будет равна, например, 85%. Эта вероятность 85% представляет собой вероят-

ность события "команда одержала победу" при условии наступления события "команда выигрывает после первого тайма". Вероятность выигрыша при условии, что команда *проигрывает* после первого тайма, будет меньше, чем общая вероятность победы, составляющая 70%; пусть, например, эта вероятность оценивается в 35%. Данная величина представляет собой вероятность события "команда одержала победу" при условии наступления события "команда проигрывает после первого тайма".

2. На успех нового коммерческого проекта влияет много факторов. Для того чтобы описать их действие, можно рассмотреть влияние на условную вероятность успеха различных факторов, таких как благоприятные или неблагоприятные экономические условия и действия конкурентов. Экономический рост будет повышать шансы на успех; это означает, что вероятность успеха при условии экономического роста будет больше, чем общая (безусловная) вероятность успеха.

Правило вычисления условной вероятности при наличии дополнительной информации

Для нахождения условной вероятности события "успех" при условии наступления события "экономический рост" необходимо вычислить, какая часть сценариев события "экономический рост" будет соответствовать окончательному результату "успех". Эта величина равна результату деления вероятности события "успех и экономический рост" на вероятность события "экономический рост".

Описанное выше соответствует общему правилу вычисления условных вероятностей. Вероятность (условная) события A при условии события B (т.е. при условии, что событие B наступило), если вероятность события B положительна, вычисляется следующим образом.[10]

Условная вероятность

$$\text{Вероятность A при условии B} = \frac{\text{Вероятность "A и B"}}{\text{Вероятность B}}.$$

Следует различать вероятность A при условии B (которая описывает вероятность события A, вычисленную с учетом того, что имеет место событие B) и вероятность B при условии A (которая описывает существенно иную ситуацию — вероятность наступления события B, вычисленную с учетом того, что имеет место событие A). Для полноты картины приведем здесь и формулу для вычисления вероятности события B при условии события A.

$$\text{Вероятность B при условии A} = \frac{\text{Вероятность "A и B"}}{\text{Вероятность A}}.$$

[10] Поскольку здесь присутствует деление на вероятность B, приведенная формула не работает в случае, если вероятность события B равна 0. В таком частном случае условная вероятность не будет определена. На практике это не создает проблем, поскольку событие, вероятность которого равна 0 (практически), никогда не наблюдается; таким образом, то, что происходит "при условии" наблюдения события B, просто не имеет значения.

Например, если вероятность того, что в неисправном приборе перегорел предохранитель, составляет 6%, вероятность обрыва провода равна 4%, а вероятность наличия обеих этих неисправностей оказывается равной 1%, можно рассчитать условную вероятность поломки провода при условии того, что в приборе перегорел предохранитель:

$$\text{Условная вероятность поломки провода при перегоревшем предохранителе} = \frac{\text{Вероятность "поломка провода и перегорел предохранитель"}}{\text{Вероятность перегорания предохранителя}} = \frac{0,01}{0,06} = 0,167.$$

В этом случае перегорание предохранителя означает повышение вероятности того, что в неисправном приборе присутствует также и обрыв провода.

Такая условная вероятность свидетельствует о том, что из всех приборов, в которых сгорел предохранитель, 16,7% обычно имеют еще и обрыв провода. Обратите внимание на то, насколько эта условная вероятность больше, чем безусловная вероятность обрыва провода (4%). Это связано с тем, что при рассмотрении приборов со сгоревшим предохранителем больше не идет речь обо "всех приборах"; теперь интерес представляют только очень немногие из них, а именно 6% всех приборов. Вероятность "обрыва провода" при этом возрастает с 4 до 16,7%, что и отражает учет этой дополнительной информации.

На рис. 6.4.7 показана диаграмма Венна для рассматриваемого случая. Обратите внимание на то, что безусловные вероятности внутри каждого из кругов представлены корректными исходными значениями (0,06 для сгоревшего предохранителя и 0,04 для поломки провода). Поскольку известно, что в приборе перегорел предохранитель, то при рассмотрении условной вероятности именно соответствующий этому событию круг становится новым выборочным пространством (ввиду того, что других возможных результатов нет). В этом новом выборочном пространстве все существовавшие ранее вероятности необходимо делить на 0,06 (безусловная вероятность того, что сгорел предохранитель), поскольку теперь результаты этого события на 100% представляют новую ситуацию.

Условные вероятности для несовместимых событий

Поскольку два несовместимых события не могут наблюдаться вместе, из информации о том, что одно из них произошло, следует, что второе *не* произошло. Это означает, что условная вероятность первого события при условии наступления второго равна нулю, конечно, если вероятность второго события не равна нулю.

Условная вероятность для двух несовместимых событий

$$\text{Вероятность A при условии B} = 0,$$

если вероятность события B не равна 0.

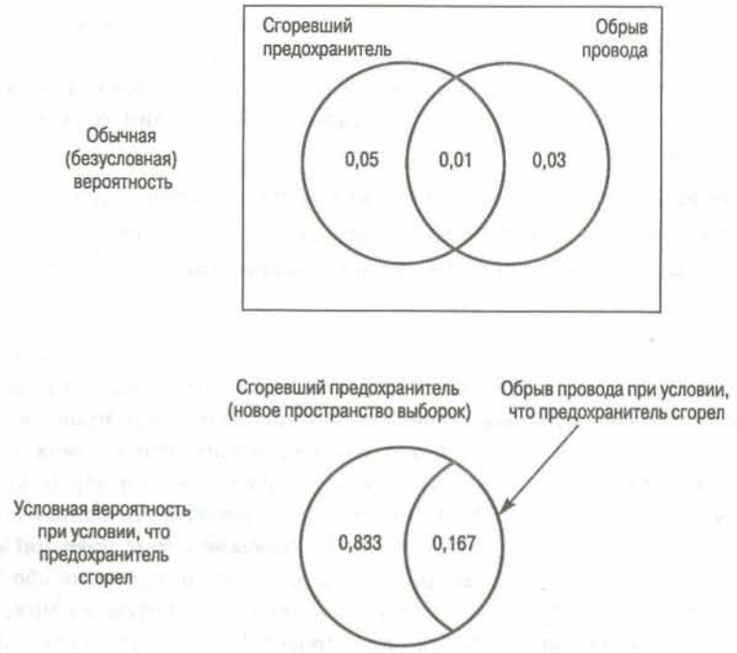


Рис. 6.4.7. Диаграмма Венна для безусловной вероятности (вверху) и условной вероятности при условии сгоревшего предохранителя (внизу). При наличии этой информации имеет смысл рассматривать только круг, соответствующий событию "предохранитель сгорел"; таким образом, этот круг становится новым полным выборочным пространством. Деление исходной вероятности на 0,06 (вероятность перегорания предохранителя) дает значения условных вероятностей в новом выборочном пространстве

Независимые события

Два события называются **независимыми**, если информация об одном из них не изменяет существовавшую до получения этой информации вероятность другого события. Если информация о некотором событии *изменяет* оценку вероятности другого события, такие события называются **зависимыми**. Например, события "быть курильщиком" и "заболеть раком" оказываются зависимыми, поскольку известно, что курильщики чаще заболевают раком, чем некурящие. С другой стороны, события "содержащиеся в вашем портфеле инвестиций ценные бумаги завтра вырастут в цене" и "завтра утром вы проспите" являются независимыми, поскольку тот факт, что вы проспите, никак не повлияет на цены фондового рынка.[11]

Формально независимость событий можно описать таким образом: события A и B являются независимыми, если вероятность A равна условной вероятности A при условии B.

[11] Если, естественно, вы – не такой влиятельный участник фондового рынка, что пропуск вами важной встречи за завтраком может сказаться на его деятельности.

События A и B независимы, если

Вероятность A = Вероятность A при условии B.

События A и B зависимы, если

Вероятность A ≠ Вероятность A при условии B.

Существует несколько способов определения того, являются ли два события независимыми. При этом, как правило, необходимо применять соответствующие формулы; "чисто умозрительные" рассуждения на тему о том, *должны* ли события быть независимыми или зависимыми, можно использовать только в качестве последнего средства в случае, когда информации для использования формул оказывается недостаточно. Ниже приведены три формулы. Использовать следует формулу, самую подходящую для конкретной имеющейся информации, поскольку все три формулы должны (в соответствии с алгебраическими правилами) всегда давать одинаковые результаты.[12]

События A и B независимы, если выполняется одно из следующих соотношений:

Вероятность A = Вероятность A при условии B;

Вероятность B = Вероятность B при условии A;

Вероятность "A и B" = Вероятность A × Вероятность B.

Третья формула позволяет найти вероятность "A и B" для случая двух событий, о которых известно, что они независимы. Однако в случае зависимых событий эта формула дает неверный результат.

Пример. Женщины на руководящих должностях

В корпорациях Сиэтла, насчитывающих 500 и более сотрудников, работает 468 руководителей. Из них 30 руководителей — женщины.[13] Если воспользоваться подходом к определению вероятности на основе относительной частоты, можно сказать, что условная вероятность того, что женщина работает на руководящей должности, составляет 30/468 = 0,064 (т.е. 6,4% женщин занимают руководящие должности). В целом среди всего населения женщины встречаются с вероятностью (безусловной) 51,2%.[14] Поскольку вероятность того, что конкретный человек принадлежит к числу женщин, изменяется при учете дополнительной информации о "работе на руководящей должности" от 51,2% до всего лишь 6,4%, данные события не являются независимыми. Это означает, что события "быть женщиной" и "занимать руководящую должность" оказываются зависимыми. Обратите внимание на то, что полученный вывод следует из приведенных выше правил (описанных соответствующими уравнениями) и чисел, а не просто из общих рассуждений по исследуемому вопросу.

[12] Следует, однако, учитывать одну техническую сложность. Если вероятность одного из событий равна 0, рассчитать условную вероятность второго события невозможно. Если одно (или оба) события имеют нулевую вероятность, мы будем (автоматически) считать их независимыми.

[13] Сведения взяты из "Seattle Corporations with 500 or More Employees", Pacific Northwest Executive, April 1988, p 20.

[14] Эта величина основывается на оценке состава населения США на 1995 год, в котором количество мужчин составляет 128685000, а количество женщин — 134749000. Данные взяты из "U.S. Bureau of the Census, Statistical Abstract of the United States: 1994" (114th ed.), Washington, D.C., 1994, p. 13.

Тот факт, что данные события зависимы, отражает исторические тенденции неравноправия полов для данной местности и данного времени: мужчинам в большей степени свойственно быть руководителями, чем женщинам (вероятности стать руководителем для мужчин и женщин различаются), а руководители чаще оказываются мужчинами, чем жители этой страны в целом (вероятности быть мужчиной для руководителя и для жителя страны различаются).

Такое исследование вероятности показывает, что неравенство полов существует, однако при этом не проясняет его причины. Если посмотреть на выраженные в процентах количественные результаты, ясно видно, что здесь есть зависимость, указывающая на существование неравенства полов. Такие различия могут быть связаны с дискриминацией при приеме на работу, с ограничениями на квалификацию для кандидатов на соответствующую должность или объясняться некоторыми другими причинами; анализ вероятностей сам по себе не указывает на то, какое из объяснений является действительно верным.

Пример. *Рыночная эффективность*

О финансовых рынках говорят, что они *эффективны*, если текущие цены отражают всю имеющуюся в наличии информацию. В соответствии с теорией рыночной эффективности невозможно получить дополнительную прибыль на основе анализа ценовой информации за предыдущий период, поскольку эта информация уже отражена в существующих ценах. Другой вывод состоит в том, что цены должны изменяться случайным образом, поскольку любые систематические изменения рынок учитывает.

Один из способов проверки эффективности рынка состоит в том, чтобы проследить существование связи между изменениями цен вчера и сегодня. Если два события "цена росла вчера" и "цена растет сегодня" независимы, это будет подтверждением эффективности рынка. В таком случае знание вчерашних ценовых тенденций не помогает в предсказании тенденций, существующих на рынке сегодня.

В то же время, если эти события зависимы, то рынок неэффективен. Так, например, если рынок обладает определенной "инерцией" и проявляет тенденцию к продолжению роста или снижению цен, сведения о вчерашнем повышении на рынке делают более правдоподобным сегодняшний рост цен. Однако подобное утверждение несовместимо с теорией рыночной эффективности, в соответствии с которой рынок учтет заранее такой дальнейший рост и различия между безусловной и условной вероятностями наблюдаться не будут.

Правило пересечения (*и*) для независимых событий

Как уже упоминалось ранее, для независимых событий (и только для независимых событий) вероятность события "A и B" можно найти простым умножением вероятностей двух рассматриваемых событий.

Правило пересечения (*и*) для независимых событий

Вероятность "A и B" = Вероятность A × Вероятность B.

Пример. *Оценка риска для большой электростанции*

Поскольку на крупных электростанциях возможно возникновение аварий, они представляют для окружающей среды и населения определенную потенциальную опасность. Несмотря на то что такая потенциальная опасность очень мала, средства массовой информации время от времени напоминают нам о том, что аварии все-таки происходят. Предположим, что вероятность перегрева на некоторой электростанции составляет 0,001 (единица к тысяче) для одного дня, а вероятность отказа резервной системы охлаждения равна 0,000001 (единица на миллион). Если предположить, что эти события независимы, вероятность "крупной аварии" (т.е. наступления события "перегрев и отказ резервной системы охлаждения") составит 0,001 × 0,000001 = 0,000000001 (единица на миллиард), что часто считается приемлемо малой вероятностью.

Однако предположение о том, что эти события независимы, может оказаться и не соответствующим истине. Может казаться, что непосредственной связи между отказом одной системы (что приводит к перегреву) и отказом другой (в результате чего система лишается резервного охлаждения) нет. Однако независимость не определяется субъективными оценками; для того, чтобы сделать вывод о независимости событий, необходимо исследовать сами вероятности. При этом вполне может оказаться, что рассматриваемые события — зависимые; так, например, может произойти природная катастрофа (наводнение или землетрясение), приводящая к выходу из строя обеих систем. Если рассматриваемые события не независимы, оценка "единица на миллиард" вероятности возникновения крупной аварии будет неверной и реальная вероятность будет намного больше.

Связь между независимыми и несовместимыми событиями

Необходимо четко различать *независимые* и *несовместимые* события. Два независимых события *не могут* быть несовместимыми (за исключением случая, когда одно из них имеет нулевую вероятность). В свою очередь, два несовместимых события не могут оказаться независимыми (опять же, за исключением случая, когда вероятность одного из них равна нулю). Если вероятность одного из событий (или обоих событий) равна нулю, события являются и независимыми, *и* несовместимыми.

6.5. Как решать вероятностные задачи

Существуют два способа решения задач на нахождение вероятности: простой и сложный. Сложный способ состоит в творческом применении правильной комбинации изложенных ранее правил, простой способ — в том, чтобы построить *дерево вероятностей* и найти ответы прямо на таком большом рисунке. Другой полезный метод, помогающий разобраться в ситуации, связан с построением *таблиц совместных вероятностей*. И еще один метод, который мы уже рассмотрели, заключается в использовании диаграмм Венна. Независимо от выбора способа — простого или сложного — ответ должен быть один и тот же.

Дерево вероятностей

Дерево вероятностей — это рисунок, на котором показаны безусловные и условные вероятности для комбинаций двух и более событий. Рассмотрим сначала пример, для которого дерево вероятностей уже построено, и проследим детали его построения. Дерево вероятностей тесно связано с *деревом решений*, которое широко используют в финансах и других областях коммерческой деятельности.

Пример. *Управление поддержкой программного обеспечения*

Поддержка программного обеспечения — достаточно сложный вид деятельности. Некоторые пользователи звонят, чтобы попросить совета, как работать с программой. Другим необходимо помочь разрешить проблемы, с которыми они столкнулись во время работы. Представьте себе, что в качестве руководителя отдела поддержки вы количественно описали вероятности некоторых характерных звонков пользователей и изобразили свои результаты в виде дерева вероятностей, показанного на рис. 6.5.1.

Рис. 6.5.1 содержит много информации. Будем рассматривать его слева направо. Прежде всего отметим, что вероятность события "пользователь раздражен" составляет 0,20 (это значит, что 20% всех обратившихся за помощью были раздражены, а 80% — нет).

Условные вероятности записаны на рисунке над четырьмя ветвями дерева, расположенными прямо под надписью "Получил ли пользователь помощь?". Обратите внимание, что 15% раздраженных пользователей помощь получили (это вероятность события "пользователь получил помощь" при условии события "пользователь раздражен"), а 85% раздраженных пользователей помощи не получили. Ниже нарисованы две другие ветви. Они свидетельствуют о том, что помощь получили 70% "нераздраженных" пользователей и не получили помощь 30% таких пользователей. Явно видно, что отдел поддержки лучше справляется с оказанием помощи пользователям, которые при обращении не высказывают своего раздражения (соотношение получивших помощь пользователей для этих групп составляет 70% к 15%).

Числа в кружках в правой части рис. 6.5.1 показывают вероятности различных событий, сформированных путем комбинирования *и* и *не*. Вероятность того, что пользователь был раздражен и получил помощь, составляет 0,03. Это означает, что 3% всех пользователей были раздражены и получили при этом помощь. Далее, 17% всех пользователей были раздражены и помощи не получили; в 56% случаев пользователи не были раздражены и помощь получили, а 24% пользователей не были раздражены и помощи не получили.

Из представленного на рисунке дерева можно найти любую из представляющих интерес вероятностей. Вероятность события "пользователь раздражен" приведена в первом столбце обведенных в кружок чисел (0,20), а вероятность противоположного события, "не раздражен", показана в кружке, расположенном непосредственно ниже. Вероятность события "пользователь получил помощь" находим сложением двух помещенных в кружочках справа вероятностей, характеризующих получение помощи: 0,03 + 0,56 = 0,59. Условная вероятность события "пользователь получил помощь" при условии наступления события "пользователь раздражен" приведена над соответствующей ветвью дерева, она равна 0,15. Несколько сложнее найти вероятность события "пользователь раздражен" при условии события "пользователь получил помощь". Для этого можно воспользоваться определением условной вероятности.[15]

$$\text{Условная вероятность "пользователь раздражен" при условии "пользователь получил помощь"} = \frac{\text{Вероятность "пользователь раздражен и получил помощь"}}{\text{Вероятность "пользователь получил помощь"}}$$

$$= \frac{0,03}{0,03+0,56} = 0,051.$$

Таким образом, из всех пользователей, которым сотрудники отдела оказали помощь, раздраженными при обращении были 5,1%. Другой способ найти эту условную вероятность состоит в том, чтобы построить новое дерево вероятностей, начинающееся не с события "пользователь раздражен", а с события "пользователь получил помощь", поскольку для того, чтобы представить некоторую условную вероятность, информацию, задающую условие, необходимо разместить в дереве перед этой условной вероятностью.

Правила построения дерева вероятностей

Для построения дерева вероятностей прежде всего необходимо нарисовать само дерево, затем записать на рисунке всю известную для данной задачи информацию и, наконец, воспользоваться основными правилами, чтобы вычислить недостающие числа и закончить дерево.

1. Вероятности указываются в каждой из конечных точек и обводятся кружочками. На каждом уровне дерева сумма этих вероятностей должна равняться 1 (или 100%). Так, например, на рис. 6.5.1 сумма вероятностей на первом уровне составляет 0,20 + 0,80 = 1,00 и на втором уровне — 0,03 + 0,17 + 0,56 + 0,24 = 1,00. Это правило помогает заполнить один пустой кружок в столбце, если значения всех остальных вероятностей этого уровня известны.

[15] Это единственная формула из предыдущего рассмотрения, которая может понадобиться. Все остальное дерево вероятности сделает само!

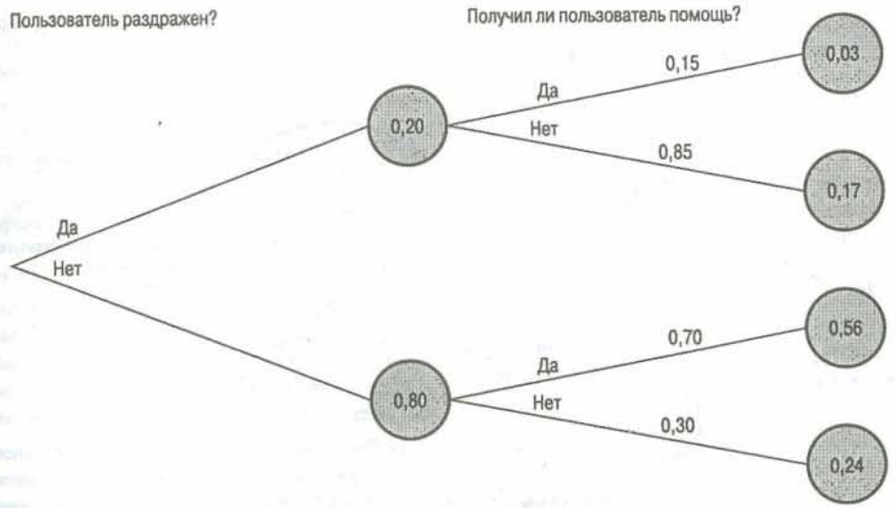


Рис. 6.5.1. Дерево вероятностей для событий "пользователь раздражен" и "пользователь получил помощь". Цифрами в кружочках показаны вероятности; остальные числа обозначают величины условных вероятностей

2. Условные вероятности указываются рядом с каждой из ветвей (кроме, возможно, ветвей первого уровня). Для каждой из групп ветвей, выходящих из одной точки, сумма этих вероятностей также равна 1 (или 100%). Например, на рис. 6.5.1 для первой группы ветвей получаем 0,15 + 0,85 = 1,00 и для второй группы — 0,70 + 0,30 = 1,00. Это правило позволяет вычислить одно неизвестное значение условной вероятности в группе ветвей, исходящих из одной точки.
3. Обведенная кругом в начале ветви вероятность, умноженная на условную вероятность рядом с этой ветвью, дает вероятность, записанную в круге в конце ветви. Например, на рис. 6.5.1 для верхней ведущей вправо ветви имеем 0,20 × 0,15 = 0,03, для следующей ветви — 0,20 × 0,85 = 0,17; аналогичные соотношения выполняются и для других двух ветвей. Это правило можно использовать для вычисления одного неизвестного значения вероятности из трех, соответствующих некоторой ветви.
4. Записанное в круге значение вероятности равно сумме обведенных кружками вероятностей на концах всех ветвей, выходящих из этого круга вправо. Так, например, для рис. 6.5.1 из круга со значением 0,20 выходят две ветви, на концах которых находятся обведенные кружками вероятности, сумма которых равна этому значению: 0,03 + 0,17 = 0,20. Это правило позволяет найти одно неизвестное значение вероятности в группе, включающей эту вероятность и все вероятности на концах ветвей дерева, выходящих из соответствующего круга.

Используя эти правила можно, зная все, кроме одного значения вероятности для некоторой ветви или на некотором уровне, находить это неизвестное значение. Общий вид такого типа дерева вероятностей с обозначением смысла всех

помечающих дерево чисел приведен на рис. 6.5.2,а. Общий вид соответствующей диаграммы Венна показан для сравнения на рис. 6.5.2,б.

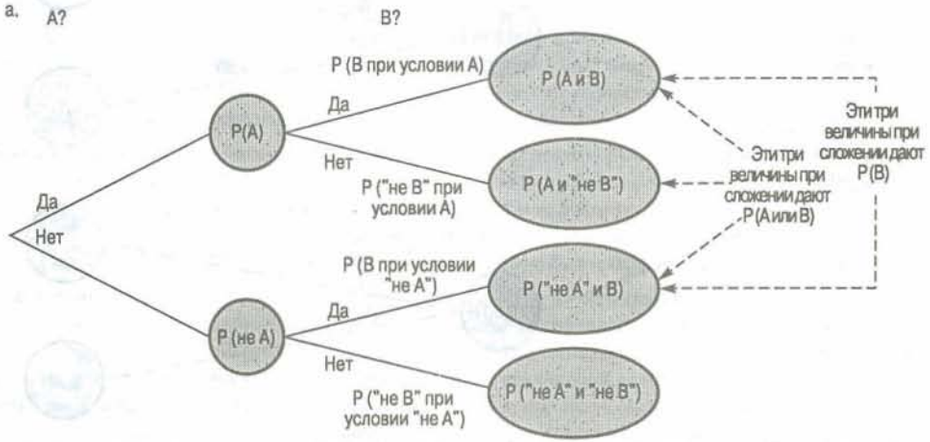


Рис. 6.5.2. а) Дерево вероятностей с показанными на нем вероятностями (в кругах и овалах) и условными вероятностями (рядом с ветвями); б) Диаграмма Венна, включающая четыре базовые вероятности. Эти четыре вероятности соответствуют вероятностям, размещенным справа на концах ветвей дерева вероятностей

Пример. *Проверка сотрудников на употребление наркотиков*

Представьте себе, что ваша фирма собирается провести обязательную проверку всех сотрудников на предмет употребления ими наркотиков. Чтобы оценить требуемые затраты (необходимые для тестирования средства и возможные психологические проблемы) и ожидаемый выигрыш (повышение производительности труда), принято решение исследовать различные результаты на основе рассмотрения ситуации для одного работника. Вы собираетесь воспользоваться деревом вероятностей для вычисления недостающей, но полезной информации.

Процедура тестирования неидеальна. Сотрудники лаборатории сообщили, что если человек употребляет наркотики, тест будет "положительным" с вероятностью 90%. Вместе с тем, если человек наркотики не

употребляет, тест покажет "отрицательный" (т.е. "не положительный") результат в 95% случаев. На основе неофициального опроса некоторых рабочих можно ожидать, что примерно 8% всего персонала употребляют наркотики.

Базовое дерево вероятностей для данной ситуации показано на рис. 6.5.3. Событие "употребляет наркотики" помещено на нем первым, поскольку часть исходной информации представлена как условная вероятность, для которой это событие выступает условием.

После нанесения на диаграмму исходной информации получаем дерево вероятностей, приведенное на рис. 6.5.4. Обратите внимание на то, что величины 90% и 95% отражают условные вероятности вдоль соответствующих ветвей; значение 8% для тех, кто употребляет наркотики, представляет собой безусловную вероятность.

Для заполнения дерева в качестве исходной информации можно использовать и значения других вероятностей, которые не всегда удается разместить непосредственно на дереве. Так, например, если бы нам сообщили вероятность (безусловную) для результата "тест положителен", ее нельзя было бы непосредственно нанести на рисунок; нужно было бы сделать примечание, что сумма значений в первом и третьем кружках на правом краю равна значению этой вероятности.

Воспользуемся теперь нашими основными правилами для нанесения недостающих значений на дерево вероятностей. Этот процесс похож на разгадывание головоломки, и идти здесь к правильному результату можно разными путями. Так, например, можно воспользоваться первым правилом, чтобы найти, что величину 0,08 дополняет величина 0,92. Второе правило дает значения условных вероятностей, 0,10 и 0,05. И, наконец, воспользовавшись третьим правилом, получаем все величины условных вероятностей для заполнения кружков в правой части. Окончательно заполненное дерево вероятностей показано на рис. 6.5.5.

Теперь легко можно построить и диаграмму Венна, воспользовавшись для этого результатами из правой части показанного на рис. 6.5.5 дерева вероятностей. Несмотря на то что диаграмма Венна для получения результатов не нужна, ее использование также часто может быть полезным.

Из дерева вероятностей (рис. 6.5.5) или диаграммы Венна (рис. 6.5.6) легко найти любую вероятность или условную вероятность. Вот несколько примеров.

Вероятность "употребляет наркотики и тест положителен" = 0,072.

Вероятность "наркотики не употребляет, но тест положителен" = 0,046.

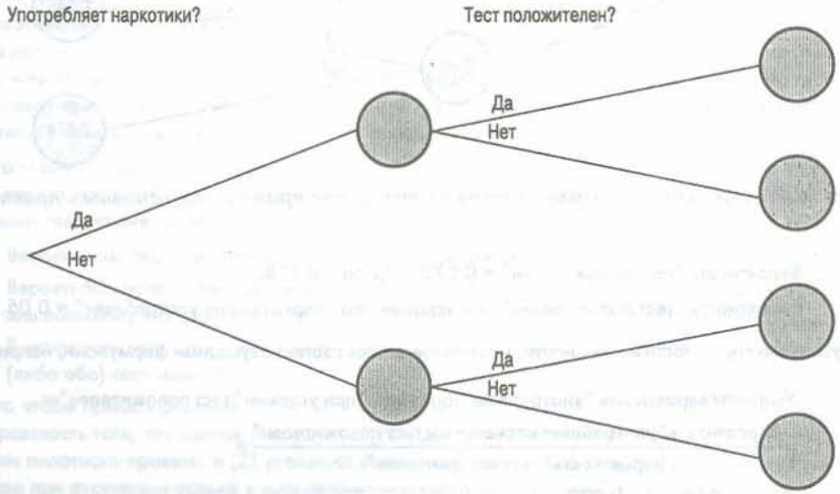


Рис. 6.5.3. Дерево вероятностей для событий "употребляет наркотики" и "тест положителен" до нанесения на него исходной информации

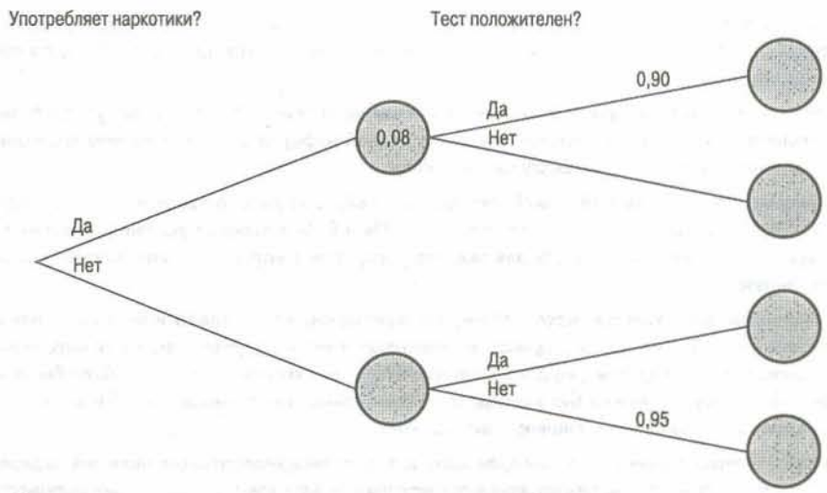


Рис. 6.5.4. Дерево вероятностей после нанесения на него исходной информации

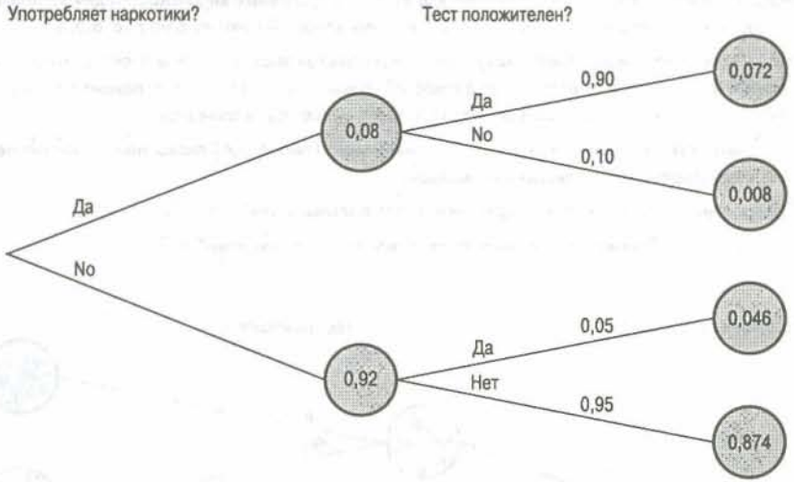


Рис. 6.5.5. Заполненное дерево вероятностей после применения основных правил

Вероятность "тест положителен" = 0,072 + 0,046 = 0,118.

Вероятность "тест положителен" при условии, что "наркотики не употребляет" = 0,05.

Другие условные вероятности можно найти, воспользовавшись соответствующими формулами, например:

$$\text{Условная вероятность "употребляет наркотики" при условии "тест положителен"} = \frac{\text{Вероятность "употребляет наркотики и тест положителен"}}{\text{Вероятность "тест положителен"}} =$$

$$= \frac{0,072}{0,072+0,046} = \frac{0,072}{0,118} = 0,610.$$



Рис. 6.5.6. Диаграмма Венна для примера проверки работников на употребление наркотиков с указанием четырех основных вероятностей

Эта условная вероятность представляет особый интерес. Несмотря на кажущуюся надежность методики тестирования (90% положительных результатов для людей, употребляющих наркотики, и 95% отрицательных для тех, кто их не употребляет), условная вероятность того, что в случае положительного результата теста человек действительно употребляет наркотики, оказывается равной только 61%. Это означает, что среди всех работников, для которых тест окажется положительным, только 61% окажутся действительно употребляющими наркотики, в то время как остальные 39% — нет.

Вот над этим уже необходимо поразмыслить трезво. Есть ли смысл проводить проверку, в результате которой 39% людей, для которых тестирование укажет на употребление наркотиков, окажутся обвиненными в этом напрасно? Анализ вероятностей имеет в данном случае большое значение, поскольку позволяет трансформировать имеющуюся информацию в значительно более полезные для принятия решений вероятности.

Пример. *Использование пилотного проекта для анализа возможного успеха выпуска товара на рынок*

Допустим, что ваша фирма собирается выпустить на рынок новый товар и на вас лежит ответственность за адекватную подачу описания соответствующих возможностей высшему руководству. В данном случае существуют два основных вопроса: во-первых, следует ли вообще работать над данным проектом и, во-вторых, есть ли смысл сначала сделать вложение в *пилотный проект* для проработки его на тестируемом рынке: такой проект потребует меньших затрат и даст возможность получить некоторую информацию относительно того, велика ли вероятность успеха нового товара.

Для того чтобы собраться с мыслями, следует продумать сценарий вида что, *если*, включив в него как пилотный проект, так и собственно выпуск продукции на рынок. Предположим, что при этом представляются разумными следующие значения вероятностей.

1. Вероятность того, что выпуск товара на рынок окажется успешным, составляет 0,60.
2. Вероятность успешного выполнения пилотного проекта равна 0,70. (Это значение несколько выше, поскольку в случае пилотного проекта рынок оказывается более восприимчивым.)
3. Вероятность того, что успешным окажется либо пилотный проект, либо выпуск товара на рынок (либо оба) составляет 0,75.

Для того чтобы принять решение о том, есть ли смысл делать пилотный проект, следует найти (1) условную вероятность того, что выпуск товара на рынок будет успешным при условии достижения успеха в выполнении пилотного проекта, и (2) условную вероятность того, что выпуск товара на рынок будет успешным даже при отсутствии успеха в выполнении пилотного проекта. Кроме того, для полноты картины понадобятся также (3) вероятность того, что успешными будут и выпуск товара на рынок, и выполнение пилотного проекта, а также (4) вероятность провала обоих.

Все перечисленные вероятности можно найти, творчески применив для этих целей основные формулы, приведенные в разделе 6.4. Однако поиск правильной для данного конкретного случая комбинации формул может потребовать значительных затрат времени. Более простой путь вычисления необходимых *значений вероятностей состоит* в построении дерева вероятностей, которое и поможет найти ответы на поставленные вопросы.

На рис. 6.5.7 показано базовое дерево вероятностей с представленной на нем исходной информацией. Обратите внимание на то, что для двух из трех известных чисел непосредственно на дереве вероятностей места нет; их можно указать рядом, как это показано на рисунке.

Что делать дальше? Первое правило построения дерева вероятностей (или правило дополнительности событий) дает возможность найти вероятность дополнительного, или противоположного, события, величину которой следует поместить в левый нижний круг на диаграмме: 1,00 – 0,70 = 0,30. Кружочки, расположенные в правом столбце, заполнить несколько сложнее. Если три верхних значения при сложении дают 0,75, а сумма двух из них составляет 0,60, третья величина должна быть равна разности этих чисел, т.е. 0,75 – 0,60 = 0,15. Это значение записываем во второй сверху кружок (вероятность "успешный пилотный проект и неуспешный выпуск товара"). Теперь можно воспользоваться правилом 4 для нахождения вероятности, которую следует поместить в верхний круг: 0,70 – 0,15 = 0,55. Итак, у нас заполнены два из тех верхних кружочков в правом столбце; поскольку все они в *сумме дают величину 0,75, неизвестная* до сих пор величина равна 0,75 – 0,55 – 0,15 = 0,05. С этого момента для заполнения дерева вероятностей можно использовать основные правила. Результат в виде заполненного дерева вероятностей показан на рис. 6.5.8.

В полностью заполненном дереве вероятностей можно найти значения всех необходимых вероятностей (*а также значения любых других вероятностей*, которые также могут представлять интерес). Ниже эти вероятности перечислены вместе с краткими комментариями к интерпретации условных вероятностей.

1. Вероятность успешного выпуска товара на рынок при условии успешного выполнения пилотного проекта составляет 0,786. Если бы пилотный проект был идеальным средством прогноза успеха новой продукции, эта величина была бы равна 1,00. Однако в реальной жизни ситуация неидеальна и после успешного выполнения пилотного проекта шансы нового товара на успех оцениваются только в 78,6%.
2. Вероятность того, что выпуск товара на рынок будет успешным в случае провала пилотного проекта равна 0,167. Если бы пилотный проект мог полностью предсказать наличие или отсутствие успеха в дальнейшем, эта величина была бы равна 0. Однако в этом случае мы видим, что с вероятностью 16,7% выпуск товара может быть успешным даже после провала пилотного проекта.

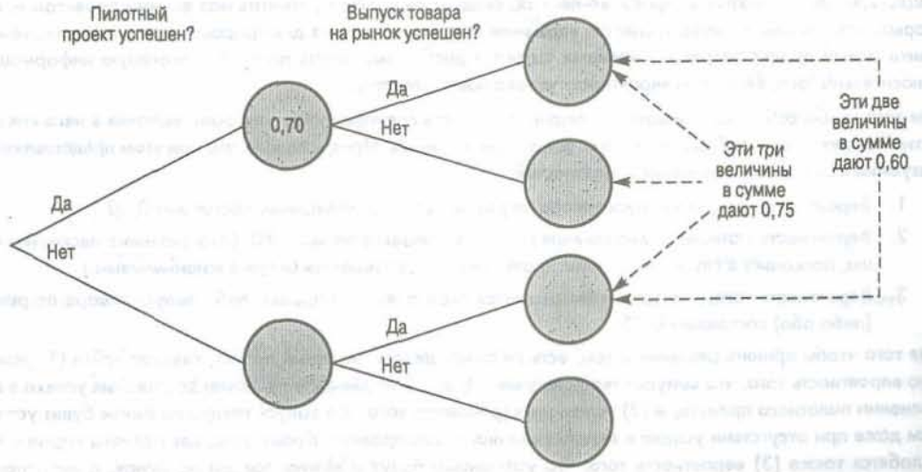


Рис. 6.5.7. Исходное дерево вероятностей до применения основных правил

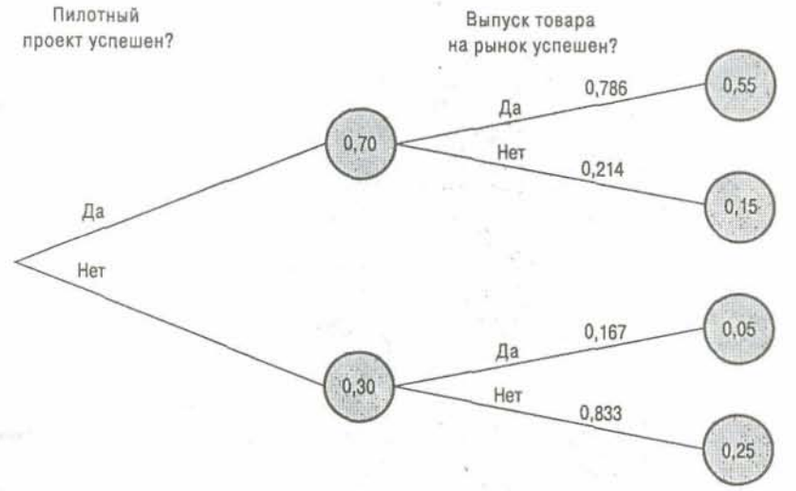


Рис. 6.5.8. Заполненное дерево вероятностей с применением основных правил

3. Вероятность того, что успешными будут и выполнение пилотного проекта, и выпуск товара на рынок, составляет 0,55.
4. Вероятность провала обоих составляет 0,25.

Пример. *Решение задачи "За какой из дверей спрятан приз?"*

Теперь мы можем построить дерево вероятности и для примера, приведенного в разделе 6.1. Разумно предположить, что приз может находиться за любой из этих дверей (т.е. у нас нет никаких подсказок относительно того, за какой дверью он спрятан), и выбор двери производится случайно[16]. Дерево вероятностей для данной ситуации показано на рис. 6.5.9. Это дерево несколько отличается от тех, которые мы строили ранее, поскольку из каждой вершины выходит *три* ветви. Однако основные правила для дерева вероятностей справедливы по-прежнему.

При построении дерева исходными являются вероятность того, что приз находится за определенной дверью (1/3), и условные вероятности выдвигаемых предположений (также 1/3 — условная вероятность всегда одна и та же, поскольку неизвестно, за какой из дверей находится приз). Далее, следуя правилам заполнения дерева вероятностей, приходим к четкому ответу: изменение выбора удваивает шансы на выигрыш с 1/3 до 2/3.

Таблица совместных вероятностей

Таблица совместных вероятностей для двух событий содержит значения вероятностей самих событий, противоположных им событий, а также комбинаций (с использованием *и*) событий. Ниже приведена таблица совместных вероятностей для рассмотренного в предыдущем разделе примера проверки работников на предмет употребления ими наркотиков.

[16] Предположение о случайности выбора не является необходимым. Действительно, вы можете всегда выбирать дверь №1, однако предположение о случайности выбора упрощает анализ задачи.

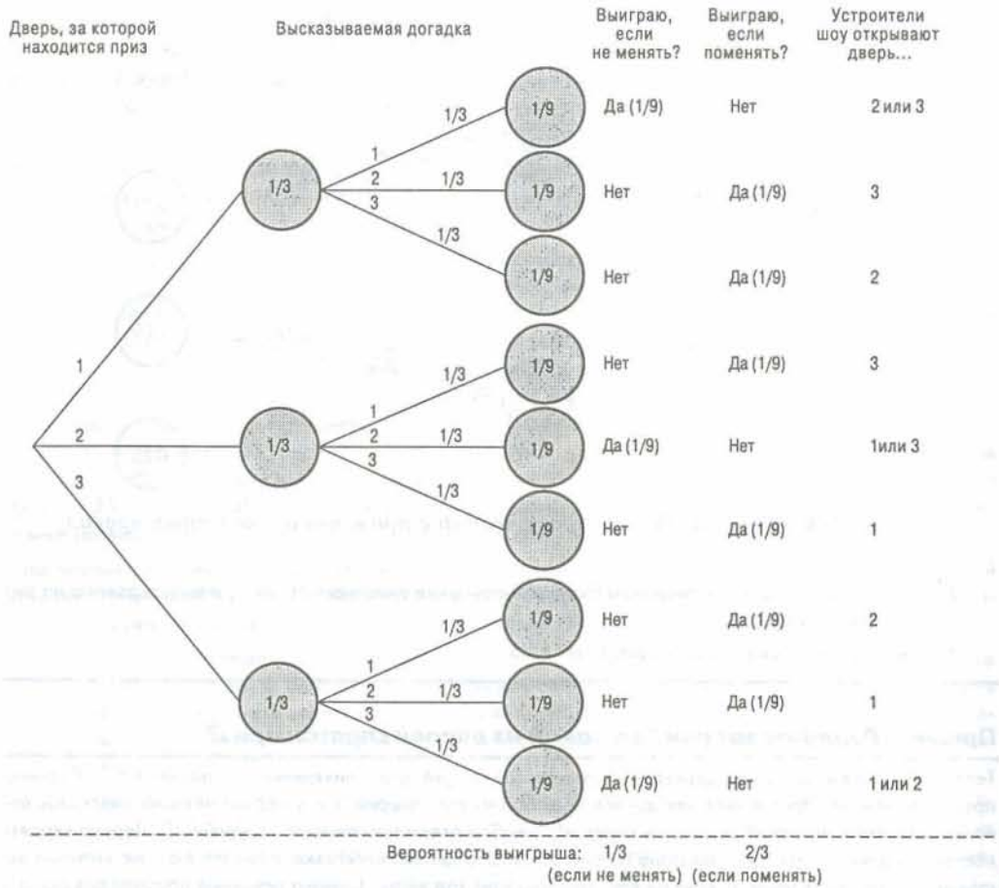


Рис. 6.5.9. Заполненное дерево вероятностей с применением основных правил

		Тест положителен?		
		Да	Нет	
Употребляет наркотики?	Да	0,072	0,008	0,08
	Нет	0,046	0,874	0,92
		0,118	0,882	1

Числа в ячейках таблицы — это четыре обведенных кругами числа в правой части дерева вероятностей. Числа вне таблицы — итоговые значения, которые называют *безусловными вероятностями* и описывают вероятности каждого из событий и его дополнения.

На рис. 6.5.10 показан общий вид таблицы совместных вероятностей и приведена интерпретация значений. Обратите внимание, что здесь нет условных вероятностей; их можно легко найти, применяя основную формулу вычисления условной вероятности.

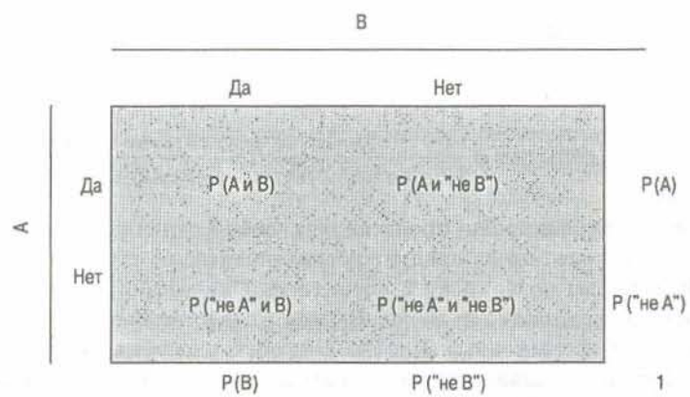


Рис. 6.5.10. Общий вид таблицы совместных вероятностей

6.6. Дополнительный материал

Резюме

Для понимания случайных, непредсказуемых ситуаций реального мира следует начинать с четкого описания существующих возможностей и тщательного построения строгой схемы исследования соответствующих вероятностей. **Случайный эксперимент** — это четко определенная процедура, результат которой можно наблюдать, но невозможно точно предсказать заранее. Каждый случайный эксперимент характеризуется **выборочным пространством**, представляющим собой набор *всех возможных результатов*. Выборочное пространство формируется заранее, когда еще неизвестно, каким окажется результат конкретного выполнения случайного эксперимента. При каждом выполнении случайного эксперимента реализуется только один **результат**, представляющий собой исход случайного эксперимента и описывающий наблюдаемые последствия данного эксперимента. Каждый раз при выполнении случайного эксперимента либо происходит либо не происходит некоторое **событие**; формально событие представляет собой некоторый определенный заранее, до проведения эксперимента, набор результатов. В каждой конкретной ситуации может быть одно или более представляющих интерес событий.

Каждому событию соответствует число от 0 до 1, называемое **вероятностью** и описывающее, насколько правдоподобно наступление данного события при каждом выполнении случайного эксперимента. Если случайный эксперимент проводится много раз, можно найти **относительную частоту** события, равную частному от деления количества раз наблюдения данного события на число выполненных случайных экспериментов. В соответствии с **законом больших чисел** при многократном повторении эксперимента относительная частота (представляющая собой случайное число) приближается к вероятности (точному, фиксированному значению). Таким образом, относительную частоту, основанную на полученных ранее данных, можно использовать в качестве приближенного значения вероят-

ности. **Теоретическая вероятность** рассчитывается с использованием точной, основанной на математической теории, формулы или модели, такой, как правило *равной вероятности*; таким образом, при равновероятных результатах

$$\text{Вероятность события} = \frac{\text{Количество результатов в событии}}{\text{Общее количество возможных результатов}}.$$

Субъективная вероятность представляет собой мнение определенного лица (если есть возможность, здесь следует воспользоваться услугами эксперта в соответствующей области) по вопросу о вероятности некоторого события. Применяемый в статистическом анализе **метод Байеса** позволяет использовать субъективные вероятности в формальных математических расчетах. Метод, альтернативный методу Байеса, называется **частотным анализом**. Этот метод не использует субъективные вероятности для расчетов, однако и он не является полностью объективным, поскольку предварительные мнения оказывают определенное влияние на выбор данных и модели (математической основы).

Диаграмма Венна представляет собой рисунок, на котором представлено множество всех возможных результатов (выборочное пространство). Она имеет вид прямоугольника, внутри которого показаны события, часто в виде кругов или овалов. Пример диаграммы Венна приведен на рис. 6.6.1. **Дополнение** события, или противоположное событие, представляет собой другое событие, которое наблюдается только в том случае, когда первое событие не происходит. В соответствии с правилом дополнительности событий

Вероятность "не А" = 1 − вероятность А.

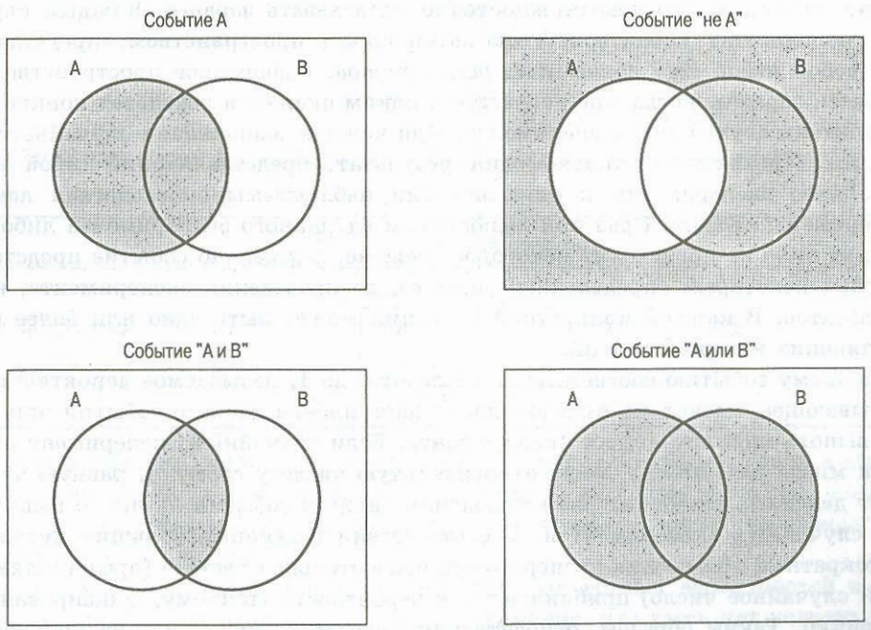


Рис. 6.6.1. Диаграмма Венна четко иллюстрирует смысл операторов не, и, *а также* или, *определяющих события через другие события*

Пересечением, или произведением, двух событий называется событие, которое наступает тогда, когда в результате однократного выполнения случайного эксперимента происходят и одно событие *и* другое.

Два события, которые не могут наступить одновременно, называются **несовместимыми событиями**. Для несовместимых событий существуют следующие правила:

Вероятность "А и В" = 0;

Вероятность "А или В" = Вероятность А + Вероятность В.

Объединение, или сумма, двух событий — это событие, которое наступает тогда, когда в результате однократного выполнения случайного эксперимента происходит или одно событие, *или* другое (или оба эти события вместе). Зная любые три из четырех вероятностей (вероятностей событий А, В, "А и В" и "А или В"), можно найти неизвестное четвертое значение с помощью одной из приведенных ниже формул.

Вероятность "А или В" =
= Вероятность А + Вероятность В − Вероятность "А и В".

Вероятность "А и В" =
= Вероятность А + Вероятность В − Вероятность "А или В".

Пересмотрев вероятность события с целью учета информации о том, что произошло некоторое другое событие, получаем **условную вероятность** данного события *при условии* наступления другого события. Обычная вероятность (до учета наступления другого события) называется **безусловной вероятностью**. Условные вероятности находят следующим образом (если вероятность составляющего условие события равна нулю, то условная вероятность не определена):

$$\text{Условная вероятность А при условии В} = \frac{\text{Вероятность "А и В"}}{\text{Вероятность В}}.$$

Условная вероятность двух несовместимых событий всегда равна нулю (если только она не оказывается неопределенной в силу технических причин).

Два события называются **независимыми событиями**, если информация об одном событии не изменяет оценку вероятности другого события. Если же информация об одном событии изменяет оценку вероятности другого, события называются **зависимыми**. Для определения того, являются ли два события зависимыми или независимыми, можно использовать одну из приведенных ниже формул. События А и В независимы, если выполняется любое из следующих соотношений:

Вероятность А = Условная вероятность А при условии В;

Вероятность В = Условная вероятность В при условии А;

Вероятность А и В = Вероятность А × Вероятность В.

Третьей формулой можно воспользоваться для вычисления вероятности "А и В" для двух событий, о которых известно, что они независимы; однако в случае зависимых событий эта формула дает неверный результат. Два независимых события не могут быть несовместимыми, за исключением случая, когда вероятность одного из них равна нулю.

Дерево вероятностей — это рисунок, на котором показаны вероятности и некоторые условные вероятности для комбинации двух или более событий. Один из самых простых способов решения задачи определения значений вероятностей состоит в построении дерева вероятностей, на котором потом можно найти требуемые ответы. Все значения вероятностей и условных вероятностей можно найти путем сложения чисел, указанных на дереве вероятностей, или путем применения формулы вычисления условной вероятности. Ниже приведены четыре правила построения и проверки правильности дерева вероятностей.

1. Вероятности указываются в каждой точке окончания ветвей дерева и обводятся кружочком. Сумма вероятностей на каждом уровне дерева должна быть равна 1 (или 100%).
2. Условные вероятности указываются рядом с каждой из ветвей (за исключением, возможно, ветвей первого уровня). Для любой группы ветвей, выходящих из одной точки, сумма соответствующих условных вероятностей равна 1 (или 100%).
3. Вероятность, записанная в кружке в начале ветви, умноженная на условную вероятность, указанную рядом с этой ветвью, равна вероятности, записанной в кружке на конце ветви (справа).
4. Записанная в кружке вероятность равна сумме обведенных кружками вероятностей на концах всех ветвей, выходящих из этого кружка вправо.

В **таблице совместных вероятностей** для двух событий приводятся вероятности событий, противоположных им событий, а также вероятности комбинаций событий, полученных с помощью операции *и*. Условные вероятности можно вычислить с помощью таблицы совместных вероятностей, воспользовавшись формулой для вычисления условной вероятности.

Основные термины

- Случайный эксперимент (random experiment), 223
- Выборочное пространство (sample space), 224
- Результат (outcome), 226
- Событие (event), 227
- Вероятность (probability), 229
- Относительная частота (relative frequency), 230
- Закон больших чисел (law of large numbers), 231
- Теоретическое значение вероятности (theoretical probability), 233
- Субъективная оценка вероятности (subjective probability), 234
- Метод Байеса (Bayesian analysis), 236
- Частотный (не Байесовский) анализ (frequentist (non Bayesian) analysis), 236
- Диаграмма Венна (Venn diagram), 237
- Дополнение (*не*) (complement (*not*)), 238
- Пересечение (*и*) (intersection (*and*)), 239
- Произведение, 239
- Несовместимые события (mutually exclusive events), 240
- Объединение (*или*) (union (*or*)), 240
- Сумма событий, 240
- Условная вероятность (conditional probability), 243
- Безусловная вероятность (unconditional probability), 261
- Независимые события (independent events), 261
- Зависимые события (dependent events), 261
- Дерево вероятностей (probability tree), 262
- Таблица совместных вероятностей (joint probability table), 262

Контрольные вопросы

1. а) Что такое случайный эксперимент?

   б) Почему определение случайного эксперимента помогает сконцентрировать внимание при исследовании неопределенной ситуации?
2. а) Что такое выборочное пространство?

   б) Есть ли какая-либо неопределенность или случайность в выборочном пространстве?
3. а) Что такое результат?

   б) Должен ли результат обязательно быть числом?
4. а) Что такое событие?

   б) Может ли в случайном эксперименте быть более чем одно представляющее интерес событие?
5. а) Что такое вероятность?

   б) Что из перечисленного можно охарактеризовать вероятностью: случайный эксперимент, выборочное пространство, событие?

   в) Если случайный эксперимент выполняется только один раз, что можно сказать о событии, вероятность которого равна 0,06?
6. а) Что такое относительная частота события?

   б) Чем относительная частота отличается от вероятности события?

   в) О чем гласит закон больших чисел?
7. а) Назовите три основных источника получения значений вероятности.

   б) Сформулируйте правило равной вероятности.

   в) Можно ли использовать в качестве значения вероятности чье-либо предположение?

   г) В чем состоит различие между анализом методом Байеса и частотным анализом?
8. Что такое несовместимые события?
9. а) Что такое дополнение события?

б) Чему равна вероятность дополнения события?

10. а) Что такое пересечение двух событий?

б) Чему равна вероятность "одно событие и другое событие", если известны следующие данные:

1) Вероятности этих событий и вероятность "одно событие или другое событие"?

2) Вероятности этих событий, а также известно, что они независимы?

3) То, что рассматриваемые события несовместимы?

11. а) Что такое объединение двух событий?

б) Чему равна вероятность "одно событие или другое", если известны:

1) Вероятности этих событий и вероятность "одно событие и другое"?

2) То, что рассматриваемые события несовместимы?

12. а) Дайте интерпретацию условной вероятности с точки зрения новой информации.

б) Всегда ли вероятность события A при условии события B выражается тем же числом, что и вероятность события B при условии события A?

в) Как можно найти условную вероятность, зная вероятности двух событий и вероятность их пересечения?

г) Чему равна условная вероятность для двух независимых событий?

д) Является ли вероятность события A при условии события B вероятностью события A или вероятностью события B?

13. а) Дайте интерпретацию независимости двух событий.

б) Как можно сделать вывод о том, являются ли два события независимыми или зависимыми?

в) При каких условиях два несовместимых события могут быть независимыми?

14. а) Что такое дерево вероятностей?

б) Сформулируйте четыре правила построения и проверки корректности дерева вероятностей.

15. Что такое таблица совместных вероятностей?

16. Что такое диаграмма Венна?

Задачи

1. Представьте себе, что вы работаете менеджером по управлению ценными бумагами в крупной маклерской конторе и по долгу службы вы должны знать все о деятельности автомобильных компаний. В частности, известно, что компания *Ford* в ближайшее время сообщит о размерах своих доходов за последний квартал, и вы не знаете, какими будут эти числа.

а) Опишите соответствующий данному случаю случайный эксперимент.

б) Что в этом случае представляет собой выборочное пространство?

в) О чем могут свидетельствовать результаты эксперимента?

г) На основе всей доступной на сегодняшний день информации вы рассчитали величину дохода в долларах и ожидаете, что рассчитанное вами значение будет достаточно близким к тому, которое будет объявлено. Точно определите событие "заявленная величина дохода выше ожидаемой" с помощью перечня результатов, содержащихся в выборочном пространстве.

д) У вас есть определенные соображения относительно вероятности того, что заявленная величина дохода будет выше ожидаемой. К какому типу вероятности относится эта вероятность, если она основана не только на предыдущем опыте, но и на вашем мнении о текущей ситуации?

2. Менеджер предприятия, выпускающего копировальные устройства, будет оценивать в конце завтрашнего рабочего дня количество произведенных устройств и количество дефектных среди них.

а) Опишите соответствующий данному случаю случайный эксперимент.

б) Каким в этом случае будет выборочное пространство?

в) О чем могут свидетельствовать результаты эксперимента?

г) Точно определите событие "достигнута цель выпуска не менее 500 исправных (не имеющих дефектов) устройств при 2 или менее дефектных устройствах" в терминах результатов, составляющих выборочное пространство.

д) В течение 22 дней за последние 25 дней эта цель достигалась. Найдите соответствующую относительную частоту.

е) Что можно сказать о том, насколько далека от истинной, неизвестной вероятности выполнения поставленной цели относительная частота, найденная в ответе на вопрос предыдущего пункта? (Для простоты можно предположить, что один и тот же случайный эксперимент выполнялся многократно и независимо.)

3. В качестве руководителя производства вы несете ответственность за составление графика работы для рабочих и для механизмов, задействованных в производственном процессе. В конце рабочего дня вы будете знать, сколько выпущено чехлов для сидений автомобиля.

а) Опишите соответствующий данному случаю случайный эксперимент.

б) Каким в этом случае будет выборочное пространство?

в) О чем могут свидетельствовать результаты эксперимента?

г) Точно определите событие "выпущено в соответствии с дневным планом, 750 единиц товара, плюс-минус 5 единиц" в терминах результатов из выборочного пространства.

д) Описанное в пункте "г" событие наблюдалось для 8 из предыдущих 15 дней. Найдите соответствующую относительную частоту.

4. Из 925 выпущенных фабрикой головных уборов 13 имеют дефекты.

а) Найдите вероятность того, что выбранный произвольным образом головной убор имеет дефекты.

б) Найдите вероятность того, что выбранный произвольным образом головной убор не имеет дефектов.

в) Какого типа эти вероятности?

5. Руководитель группы из 35 человек, работающих на ковровой фабрике, в штате которой насчитывается 118 рабочих, узнает, что в следующий понедельник из этих 118 рабочих будут выбирать одного представителя. Представитель будет выбираться случайным образом, независимо от того, входит ли он(а) в число подчиненных данного руководителя.

а) Найдите вероятность того, что представитель будет выбран из числа подчиненных данного руководителя.

б) Чему равна вероятность того, что представитель будет выбран из числа рабочих, не находящихся в подчинении этого руководителя?

6. Представьте себе, что вы отвечаете за составление графика работ строительства центра для проведения общественных мероприятий. Во избежание больших затруднений необходимо, чтобы бетон был доставлен не позднее 27 июля, а финансирование организовано до 6 августа. На основе собственного опыта и анализа аналогичных ситуаций с применением субъективной оценки вероятности вы приписываете этим двум событиям соответственно вероятности 0,83 и 0,91. Предположим также, что вероятность выполнения одного из сроков или другого (или обоих) составляет 96%.

а) Найдите вероятность возникновения "больших затруднений".

б) Являются ли данные события несовместимыми? Откуда вам это известно?

в) Являются ли данные события независимыми? Откуда это известно?

7. Два структурных подразделения работают вместе над созданием спутника связи. Для того чтобы спутник был запущен вовремя, оба подразделения должны закончить работы в срок. Вы считаете, что вероятность окончания работ в срок составляет для каждого из подразделений 90%. Если предположить, что оба эти подразделения работают независимо (так, что соответствующие события оказываются независимыми), чему равна вероятность того, что запуск спутника будет отложен по причине невыполнения сроков?

8. Вероятность получения крупного заказа, который обсуждается в настоящее время, равна 0,4. Вероятность финансовых потерь в текущем квартале составляет 0,5.

а) Предположим, что названные события несовместимы. Найдите вероятность получения заказа или потери денег.

б) Снова предположим, что рассматриваемые события несовместимы. Исключает ли это возможность того, что заказ не будет получен и не удастся заработать денег?

в) Теперь предположим, что события "получение заказа" и "финансовые потери" независимы (поскольку заказ все равно не войдет в финансовую отчетность за данный квартал). Найдите вероятность получения заказа и финансовых потерь.

г) А теперь предположим, что вероятность получения заказа и финансовых потерь составляет 0,1. Являются ли события "получение заказа" и "финансовые потери" независимыми? Из чего это следует?

9. Ваша фирма классифицирует заказы по двум признакам: как крупные или мелкие в долларовом выражении и как легкие или тяжелые в отношении отгрузочного веса. В предыдущий период 28% заказов были крупными в долларовом выражении, 13% заказов были тяжелыми, а 10% заказов были крупными в долларовом выражении и тяжелыми.

а) Постройте и заполните дерево вероятностей для данной ситуации, проведя первую ветвь для события "крупный в долларовом выражении".

б) Составьте таблицу совместных вероятностей для данной ситуации.

в) Постройте диаграмму Венна для данной ситуации.

г) Найдите вероятность того, что некоторый заказ оказывается крупным в долларовом выражении или тяжелым (или и тем и другим).

д) Найдите вероятность того, что некоторый заказ оказывается крупным в долларовом выражении, но не тяжелым.

е) Какой процент крупных в долларовом выражении заказов оказываются тяжелыми? Какой условной вероятностью это определяется?

ж) Какой процент тяжелых заказов является крупным в долларовом выражении? Какой условной вероятностью это определяется?

з) Являются ли события "крупный в долларовом выражении" и "тяжелый" несовместимыми? Из чего это следует?

и) Являются ли события "крупный в долларовом выражении" и "тяжелый" независимыми? Из чего это следует?

10. При решении вопроса о строительстве нового ресторана рассматриваются две возможности его размещения в — в южной и в северной части города. Реально только одно из этих двух мест будет доступно для застройки. Если ресторан будет построен в северной части, вероятность его успешного функционирования в течение первого года равна 90%. Если же построить ресторан в южной части, вероятность успешной работы в первый год будет составлять только 65%. Оценка вероятности того, что ресторан можно будет построить в северной части, равна 40%.

а) Постройте дерево вероятностей для данной ситуации, проведя первую ветвь для события "размещение".

б) Найдите вероятность того, что работа ресторана в первый год будет успешной.

в) Найдите вероятность того, что ресторан будет построен в южной части города и его работа будет успешной.

г) Найдите вероятность того, что ресторан будет построен в южной части города при условии, что его работа будет успешной.

д) Найдите вероятность отсутствия успеха в работе ресторана при условии того, что он построен в северной части города.

11. Следующий год ожидается удачным с вероятностью 0,70. При условии того, что год — удачный, с вероятностью 0,90 ожидается выплата дивидендов. Однако, если год окажется неудачным, выплата дивидендов произойдет с вероятностью 0,20.

а) Постройте дерево вероятностей для данного случая, выбрав в качестве первой ветви наиболее подходящее событие.

б) Найдите вероятность того, что год — удачный и дивиденды выплачиваются.

в) Найдите вероятность того, что дивиденды выплачиваются.

г) Найдите условную вероятность того, что год удачный при условии, что дивиденды выплачиваются.

12. Фирма рассматривает вопрос о выпуске новой зубной пасты. При обсуждении стратегии сделан вывод о том, что маркетинговое исследование будет удачным с вероятностью 0,65. Достигнуто также согласие по вопросу о том, что вероятность успешного выпуска товара на рынок составляет 0,40. При условии удачного маркетингового исследования вероятность успешного выпуска товара на рынок равна 0,55.

а) Постройте дерево вероятностей для данной ситуации.

б) Найдите вероятность того, что маркетинговое исследование оказывается удачным и выпуск товара на рынок также оказывается успешным.

в) При условии успешного выпуска товара на рынок найдите условную вероятность того, что маркетинговое исследование дало благоприятный результат.

г) Найдите условную вероятность того, что выпуск товара на рынок оказывается успешным при условии отсутствия успеха в маркетинговом исследовании.

д) Являются ли два события, "успешное маркетинговое исследование" и "успешный выпуск товара на рынок", независимыми? Из чего это следует?

13. Магазин заинтересован в углублении знаний о модели поведения своих покупателей и о ее связи с частотой посещений магазина. Вероятность того, что посещение магазина завершится покупкой, составляет 0,35. Вероятность того, что покупатель был в этом магазине в течение предыдущего месяца, равна 0,20. Из тех, кто ничего не купил, в последний месяц посещали магазин 12%.

а) Постройте дерево вероятностей для данной ситуации.

б) Найдите условную вероятность того, что посетитель совершит покупку при условии, что он был в магазине в течение прошлого месяца.

в) Какой процент покупателей часто посещают магазин и делают покупки, если эту категорию покупателей составляют те, кто совершает покупку и был в магазине в течение прошлого месяца?

14. Представьте себе, что вы — руководитель группы по анализу проблем контроля качества. Предположим, что вероятность дефекта формы изделия составляет 0,03, вероятность дефекта покраски равна 0,06 и эти события независимы.

а) Найдите вероятность наличия у изделия дефекта формы и дефекта покраски.

б) Найдите вероятность наличия у изделия дефекта формы или дефекта покраски.

в) Найдите вероятность того, что изделие не имеет дефектов (т.е. оба эти дефекта отсутствуют).

15. Для типичных посетителей данной торговой точки вероятность покупки бензина составляет 0,23, вероятность покупки бакалейных товаров равна 0,76, а условная вероятность покупки бакалейных товаров при условии покупки бензина равна 0,85.

а) Найдите вероятность покупки типичным посетителем и бензина, и бакалейных товаров.

б) Найдите вероятность того, что типичный посетитель совершает покупку либо бензина, либо бакалейных товаров.

в) Найдите условную вероятность покупки бензина при условии покупки бакалейных товаров.

г) Найдите условную вероятность покупки бакалейных товаров при условии, что бензин не покупается.

д) Являются ли эти события (покупка бензина, покупка бакалейных товаров) несовместимыми?

е) Являются ли эти два события независимыми?

16. Вам сообщили хорошие новости: опытный образец нового товара выпущен с опережением графика, и его функциональные качества выше, чем ожидалось. Следует ли ожидать, что "условная вероятность того, что этот товар будет иметь успех в случае хороших новостей", окажется выше, меньше или равной безусловной вероятности успеха?

17. Ваша компания рассылает заявки на участие в конкурсах для выполнения различных проектов. В тех случаях, когда вы заинтересованы выиграть конкурс (30% всех заявок) необходима большая работа по подготовке предложений; в других случаях можно ограничиться быстрыми расчетами и послать заявку, даже если вы считаете, что у этой заявки вероятность выиграть очень мала. Если в разработку заявки вкладывается много усилий, существует вероятность 80%, что в этом случае удастся заключить контракт на выполнение проекта. При подаче на рассмотрение результатов быстрых расчетов условная вероятность принятия заявки составляет только 10%.

а) Постройте дерево вероятностей для данной ситуации.

б) Чему равна вероятность того, что удастся добиться заключения контракта?

в) Если удалось заключить контракт, чему равна условная вероятность того, что в заявку было вложено много труда?

г) Если заключить контракт не удалось, чему равна условная вероятность того, что в заявку было вложено много труда?

18. В некоторой фирме 35% работников — штатные научные сотрудники, 26% — руководящие работники, а 9,1% относятся и к тем и к другим. Можно ли считать события "штатный научный сотрудник" и "руководящий работник" независимыми?

19. Отдел маркетинга некоторой фирмы провел исследование потенциальных потребителей и нашел, что (1) 27% из них читают торговое издание *Industrial Chemistry*, (2) 18% покупали товары этой фирмы и (3) 63% тех, кто читает *Industrial Chemistry*, никогда не покупали товары этой фирмы.

а) Постройте дерево вероятностей для данной ситуации.

б) Какой процент потенциальных потребителей не читают *Industrial Chemistry* и не покупали товары данной фирмы? (Эта группа отражает возможности расширения деятельности фирмы в будущем.)

в) Найдите условную вероятность того, что некоторый потребитель читает *Industrial Chemistry* при условии, что он покупал товары этой фирмы. (Это показатель распространенности публикаций среди потребителей товаров фирмы.)

20. На основе анализа данных за прошлый год было установлено, что 40% посетителей вашего магазина не бывали в нем ранее. В то время как некоторые пришли просто посмотреть, 30% посетителей что-либо купили. Однако среди тех, кто в магазине раньше не был, покупку совершили только 20%. Вы хотите использовать эти величины в качестве вероятностей для представления того, что происходит при каждом отдельном посещении магазина.

а) Какие типы вероятностей здесь указаны — с точки зрения того, каков их источник?

б) Постройте дерево вероятностей для данной ситуации.

в) Найдите вероятность того, что некоторый посетитель уже бывал в магазине раньше и совершит покупку.

г) Чему равна вероятность того, что посетитель уже бывал в магазине при условии, что в это посещение он покупки не сделал?

21. Сотрудник фирмы, отвечающий на телефонные звонки, получает много обращений по разным вопросам. В 75% случаев лишь запрашивается информация, в то время как 15% звонков связаны с реальными заказами. Кроме того, в 10% обращений запрашивается информация и делается заказ.

а) Чему равна условная вероятность того, что некоторый звонок приводит к получению заказа, если в этом же звонке еще и запрашивается информация? (Эти данные дают возможность оценить, насколько важна для бизнеса обработка запросов на получение информации.)

б) Чему равна условная вероятность того, что некоторый звонок не связан с обращением за информацией, при условии, что в результате делается заказ? (Эти данные позволяют оценить долю заказов, которые было "легко" получить.)

в) Чему равна вероятность того, что в результате обращения делается заказ и не запрашивается информация? Дайте интерпретацию.

г) Почему ответы на вопросы пунктов b и c различаются?

д) Являются ли два события, "запрошена информация" и "сделан заказ", независимыми? Из чего это следует?

22. Вы подали заявку на создание крупной коммуникационной сети. В соответствии с доступной информацией существует вероятность в 35%, что предпочтение будет отдано заявкам конкурентов. Если это произойдет, вы считаете, что все равно с вероятностью 10% вы сможете заключить контракт, найдя для этого существенную аргументацию. Однако в случае, если предпочтение будет отдано вашей заявке, существует вероятность в 5%, что вы потеряете контракт в результате действий конкурентов.

а) Постройте дерево вероятностей для данной ситуации.

б) Найдите вероятность того, что контракт удастся заключить.

в) Найдите вероятность того, что предпочтение будет отдано вашей заявке и вы сможете заключить контракт.

г) Определите условную вероятность того, что предпочтение будет отдано вашей заявке при условии, что вы заключите контракт.

д) Являются ли события "вам не удалось заключить контракт" и "предпочтение отдано вашей заявке" несовместимыми? Почему да или почему нет?

23. Вероятность успешного выполнения некоторого проекта в Нью-Йорке равна 0,6, вероятность успешного выполнения этого проекта в Чикаго составляет 0,7, а вероятность того, что данный проект будет успешным на обоих рынках, оказывается равной 0,55. Найдите условную вероятность того, что проект будет успешно выполнен в Чикаго при условии, что он успешно выполнен в Нью-Йорке.

24. Проект, связанный с кофе *экспрессо*, будет выполняться успешно с вероятностью 0,80. Вы считаете, что при условии успешного выполнения этого проекта вероятность успешного выполнения проекта, связанного с травяным чаем, составляет 0,70. Однако если проект работы с кофе не будет успешным, проект по травяному чаю пойдет хорошо с вероятностью всего лишь 25%.

а) Постройте дерево вероятностей для данной ситуации.

б) Найдите вероятность успешного выполнения проекта работы с травяным чаем.

в) Найдите вероятность успешного выполнения обоих проектов.

г) Найдите условную вероятность успешного выполнения проекта работы с кофе при условии успешного выполнения проекта по чаю. Сравните полученное значение с безусловной вероятностью для этого же события и поясните результат.

25. Фирма отслеживает реакцию людей, получивших по почте каталог. Установлено, что 4% получивших каталог заказали шапочку и 6% заказали варежки. При условии заказа шапочки 55% заказали еще и варежки.

а) Какой процент тех, кто получил каталог, заказали оба предмета?

б) Каков процент людей, получивших каталог и не заказавших ничего?

в) Чему равен среди получивших каталог процент тех, кто отказался от шапочки, но заказал варежки?

26. 24% ваших клиентов имеют высокий доход, 17% хорошо образованны. Кроме того, 12% имеют высокий доход и хорошее образование. Какой процент имеющих хорошее образование заказчиков имеет высокий доход? Какие выводы можно сделать из полученной информации относительно маркетинговых усилий, направленных сейчас на людей с хорошим образованием, несмотря на то, что вы предпочли бы ориентироваться на людей с высоким доходом?

27. Производственная линия оснащена автоматическим сканером для выявления дефектов. В последней партии товара 2% изделий имели дефекты. Если изделие имеет дефект, то сканер определяет это изделие как дефектное с вероятностью 90%. Для изделий, не имеющих дефекта, сканер определяет их как действительно бездефектные с вероятностью 90%. Найдите условную вероятность того, что изделие действительно имеет дефект, если сканер указал на наличие в нем дефекта.

28. Установлено, что 2,1% всех компакт-дисков, выпускаемых фабрикой, имеют дефекты, обусловленные используемыми материалами, а 1,3% — дефекты, связанные с ошибками людей. В предположении независимости этих событий найдите вероятность того, что компакт-диск имеет по меньшей мере один их таких дефектов.

29. Вы считаете, что график работ можно выполнить при условии, что вовремя удастся принять на работу нового менеджера, однако, несмотря на это, ситуация остается рискованной. По вашему мнению, вероятность своевременно нанять нового менеджера равна 70%. Если менеджер будет принят на работу вовремя, вероятность успеха работы составляет 80%. Если же нового менеджера вовремя найти не удастся, вероятность успеха работы составляет только 40%. Найдите вероятность успешного выполнения графика работ.

30. Сходящая с производственной линии продукция содержит 5% дефектных деталей, которые желательно выявить до отгрузки. Быстрый и не требующий больших затрат метод проверки показал, что дефект имеют 8% деталей. Известно, что из этих деталей действительно содержат дефекты 50%.

    а) Постройте дерево вероятностей для этой ситуации.

    б) Найдите вероятность того, что дефектная деталь действительно будет выявлена (т.е. условную вероятность того, что деталь определяется как дефектная при условии, что она в действительности имеет дефект).

    в) Найдите вероятность того, что деталь имеет дефект или определяется при проверке как дефектная.

    г) Являются ли события "деталь имеет дефект" и "деталь определяется как дефектная" независимыми? Из чего это следует?

    д) Может ли метод проверки оказаться полезным в случае независимости событий "деталь имеет дефект" и "деталь определяется как дефектная"? Поясните свой ответ.

31. Существует такое высказывание относительно первого выпуска акций в продажу: "Если они вам нужны, их невозможно достать; если вы можете их достать, значит, они вам не нужны". Подобное мнение связно с тем, что часто оказывается сложно получить впервые поступающие в продажу акции новой привлекательной компании. Большинству инвесторов приходится ждать начала свободной торговли такими акциями на бирже, часто по достаточно завышенным ценам. Предположим, что, при условии возможности приобрести акции первого выпуска, вероятность высокой доходности таких акций составляет 0,35. Однако если такой возможности нет, условная вероятность (при условии отсутствия возможности покупки) высокой доходности равна 0,8. Предположим также, что в целом возможность приобретения акций первого выпуска составляет для вас 15%.

    а) Постройте дерево вероятностей для данной ситуации.

    б) Найдите вероятность того, что вы можете (1) приобрести акции первого выпуска и (2) эти акции имеют высокую доходность.

    в) Насколько велик ваш доступ к имеющим высокую доходность акциям первого выпуска? Ответ на этот вопрос можно получить, найдя условную вероятность того, что вы можете купить акции первого выпуска, при условии, что эти акции имеют высокую доходность.

    г) В какой части случаев (в процентах), в течение длительного периода, вы будете удовлетворены результатом своей деятельности в области работы с акциями первого выпуска? Это означает, насколько вам удается приобретать акции с высокой доходностью, или, наоборот, не покупать акций, доходность которых впоследствии оказывается низкой.

32. Вероятность получения патента равна 0,6. Если вы получите патент, условная вероятность получения дохода от него составит 0,9. Однако если патент не будет получен, условная вероятность получения дохода составляет только 0,3. Найдите вероятность получения дохода.

33. Вы принимаете участие в телевизионном шоу и боретесь за получение приза, спрятанного за одной из пяти дверей. Есть только один приз, и он спрятанный за одной из этих дверей, выбранной случайно. После того как вы сделали свой выбор, устроители шоу сознательно открывают три двери (за исключением той, которую вы выбрали), за которыми *нет* приза. У вас есть возможность изменить свой первоначальный выбор и выбрать другую неоткрытую дверь.

    а) Чему равна вероятность получения приза, если изменить выбор?

    б) Чему равна вероятность получения приза, если выбор не менять?

34. Рассмотрим игру в казино, в котором посетители выигрывают с вероятностью 0,40. Известно, что вчера здесь играли 42652 человека, причем выиграли из них 17122.

    а) Найдите относительную частоту выигрыша и сравните ее с вероятностью.

    б) Как закон больших чисел позволяет владельцу казино, в котором играет очень много людей, в значительной степени избежать связанной с азартной игрой неопределенности?

в) Помогает ли закон больших чисел человеку, который сыграет один или два раза, уменьшить неопределенность? Почему да или почему нет?

35. Ваша новая фирма выпускает на рынок два вида товаров: прицеп к велосипеду и детскую коляску. Ваша субъективная оценка вероятности того, что эти два товара будут иметь успех на рынке, составляет соответственно 0,85 и 0,70. Если прицеп будет пользоваться успехом, у вас появится возможность продвигать на рынок коляску, предлагая ее покупателям прицепа; в связи с этим вы ожидаете, что в случае успешной работы на рынке с прицепами работа с коляской будет успешной с вероятностью 0,80.

    а) Постройте дерево вероятностей для этой ситуации.

    б) Найдите вероятность того, что работа с обоими этими видами товаров будет успешной.

    в) Найдите вероятность того, что успех не будет достигнут ни для одного из этих двух товаров.

    г) Найдите вероятность того, что торговля прицепами будет успешной, а колясками — нет.

    д) Для выживания вашей фирмы необходимо, чтобы успешной оказалась работа хотя бы с одним из этих видов товаров. Найдите вероятность выживания фирмы.

36. Когда Палата представителей подготовила к выходу в свет видеозапись показаний большому жюри Президента Клинтона, его рейтинг одобрения был таким: 36% одобряли его как личность, 63% одобряли его как президента, 30% одобряли его как президента, но не как личность.[17] Найдите процент людей, которые одобряли его как личность, но не как президента. Постройте соответствующую данному случаю диаграмму Венна.

Упражнения с использованием базы данных

Обратитесь к базе данных наемных работников, приведенной в приложении А.

1. Будем считать эту базу данных выборочным пространством некоторого случайного эксперимента, в котором случайным образом выбирается работник. Таким образом, один работник представляет один результат и все возможные результаты равновероятны.

    а) Найдите вероятность того, что будет выбрана женщина.

    б) Найдите вероятность того, что зарплата превышает $35000.

    в) Найдите вероятность того, что работник имеет уровень подготовки B.

    г) Найдите вероятность того, что зарплата превышает $35000 и работник имеет уровень подготовки B.

    д) Найдите вероятность того, что зарплата превышает $35000 при условии, что работник имеет уровень подготовки B.

[17] Основано на R. Mishra, "Swing' Group Holds the Key", The Seattle Times, September 19, 1998, p. A2. Использованы источники: газеты Knight Ridder и CNN/USA Today Gallup Poll, исследование J. Treible.

е) Является ли событие "зарплата выше $35000" независимым от события "уровень подготовки B"? Из чего это следует?

ж) Найдите вероятность того, что заработная плата превышает $35000 при условии, что работник имеет уровень подготовки C.

2. Снова, как и в предыдущей задаче, будем рассматривать базу данных работников в качестве выборочного пространства. Рассмотрим два события: "большой опыт работы (шесть лет или более)" и "работник — женщина".

    а) Найдите вероятности этих двух событий.

    б) Найдите вероятность их пересечения. О чем свидетельствует полученный результат?

    в) Постройте дерево вероятностей для этих двух событий, выбрав в качестве первой ветви "работник — женщина".

    г) Найдите условную вероятность наличия большого опыта работы при условии, что работник — женщина.

    д) Найдите условную вероятность того, что работник — женщина, при условии наличия большого опыта работы.

    е) Найдите вероятность того, что работник — мужчина, не имеющий большого опыта работы.

    ж) Являются ли события "работник — женщина" и "имеет большой опыт работы" независимыми? Из чего это следует?

    з) Являются ли события "работник — женщина" и "имеет большой опыт работы" несовместимыми? Из чего это следует?

3. Снова считаем базу данных, которую мы использовали в задаче 1, выборочным пространством.

    а) Являются ли события "уровень подготовки A" и "уровень подготовки B" независимыми? Из чего это следует?

    б) Являются ли события "уровень подготовки A" и "уровень подготовки B" несовместимыми? Из чего это следует?

Проект

Выберите некоторую связанную с вашими деловыми интересами задачу, которая требует принятия решения с вовлечением в рассмотрение двух случайных событий.

а) Выберите разумные исходные значения для трех величин вероятностей.

б) Постройте дерево вероятностей.

в) Выделите две безусловные и две условные вероятности, имеющие отношение к вашей задаче, и проинтерпретируйте их.

г) Опишите (в одном абзаце), что вы узнали о принятии решений в результате выполнения этой работы.

Ситуация для анализа

Детективная история: кто же все-таки ответствен за увеличение количества дефектов в последнее время?

Тяжелый случай. Процент дефектной продукции в последнее время резко вырос, и на вас, с целью исправления ситуации, возложена задача по выявлению проблемы. Двое из трех отвечающих за работу производственной линии менеджеров (Джонс, Уоллес и Ландуэл) к вам уже заходили (как и некоторые из рабочих). Рассказали они любопытные вещи.

Кто-то обвиняет во всем Джонса, используя слова "безответственный" и "все еще изучает азы" и приводя в подтверждение своих слов примеры ведения работ. Кое-что из рассказанного — явное следствие конкуренции между сотрудниками, и это, безусловно, необходимо учитывать, однако вопрос все же стоит рассмотреть. Джонс же ссылается на то, что процент дефектов производства выше не в его смену и утверждает, что фактически у Уоллеса выход бракованной продукции оказывается значительно выше. Свои слова Джонс подкрепляет такими данными.

	Процент дефектной продукции
Уоллес	14,35
Джонс	7,84

Вскоре после этого появляется Уоллес (который, как известно, особым тактом не отличается) с криками, что Джонс — (нецензурное определение) ... и ему верить нельзя. Несколько поостыв, Уоллес начинает что-то невнятно бормотать о том, что вышестоящие руководители дают ему сложные задания. Однако даже на прямой вопрос о проценте дефектных изделий четкий ответ от него получить не удается. Ваши подозрения крепнут: похоже, проблема действительно где-то здесь. Однако вам известно и то, что Уоллес (если абстрагироваться от его манер) на хорошем счету у технических экспертов и не следует выдвигать против него обвинений, не рассмотрев сначала возможные пояснения и альтернативные варианты.

В такой ситуации, естественно, следует определить и выход дефектной продукции в смену Ландуэлла, а также результаты по двум типам продукции: для потребителей внутри страны и для иностранных клиентов (в последнем случае спецификации должны выдерживаться значительно точнее). Такие данные образуют более полный набор, который и представлен ниже для произведенной в последнее время продукции.

	С дефектами	Без дефектов
Внутреннее потребление		
Уоллес	3	293
Ландуэлл	12	307
Джонс	131	2368
Экспорт		
Уоллес	255	1247
Ландуэлл	75	359
Джонс	81	123

Окончание таблицы

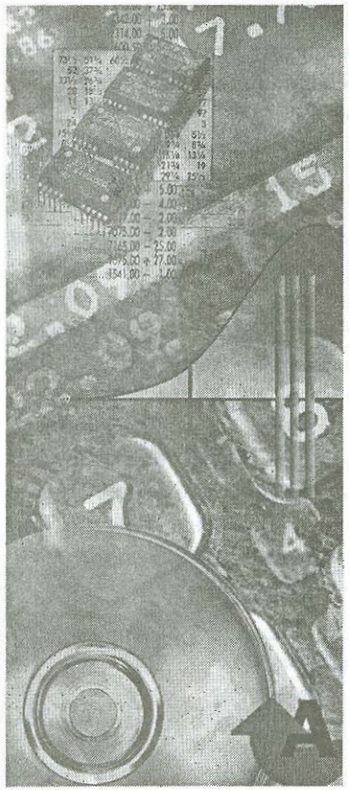
Вопросы для обсуждения

1. Прав ли Джонс? А именно: подтверждается ли при использовании более полного набора данных утверждение о том, что у Джонса выход бракованной продукции ниже всего? Верны ли представленные Джонсом процентные значения в целом (т.е. для всей продукции, как для внутреннего потребления, так и предназначенной на экспорт)?
2. Прав ли Уоллес? А именно: к какой части выпущенной в его смену продукции предъявляются повышенные требования? Какие выводы можно сделать при сравнении экспортной продукции, выпущенной в смену Уоллеса, с экспортной продукцией двух других смен? (Примечание. Возможно, для этого будет полезно сравнить условные вероятности выпуска продукции с дефектами и без дефектов при условии руководства сменой определенным менеджером.)
3. Внимательно проанализируйте условные вероятности выпуска дефектной продукции в случае различных комбинаций менеджера и заказчика товара. Какие выводы можно сделать из этих данных?
4. Стали бы вы рекомендовать Уоллесу начать поиск новой работы? Если нет, то каковы ваши предложения?

ГЛАВА 7

Случайные величины: работа с неопределенными значениями

В деловой жизни часто возникают ситуации, требующие работы со случайными величинами, — например, при определении эффективности портфеля инвестиций или проведении маркетингового опроса потребителей с целью выяснить, сколько они могут потратить на покупки. Каждый раз, когда частью результата случайного эксперимента является число (одно или несколько), мы имеем дело со случайными величинами (случайными переменными). Для нас, естественно, важно уметь вычислять и анализировать некоторые обобщающие характеристики (такие как типичное значение и риск), а также вероятности событий, которые зависят от результата наблюдения, например, вероятность того, что стоимость портфеля возрастет на 10% или больше.



Случайные величины можно также рассматривать в качестве источников данных. Многие из тех наборов данных, с которыми мы работали в главах 2–5, были получены в результате наблюдений и фиксации значений некоторых случайных величин. В этом смысле сама случайная величина уже представляет некоторую генеральную совокупность (или процесс выборки из генеральной совокупности), в то время как наблюдаемые значения случайной величины представляют собой результат выборки. В главе 8 генеральные совокупности и выборки будут рассмотрены более детально; в этой главе мы сосредоточим наше внимание на случайных значениях.

Ниже приведено несколько примеров случайных величин. Обратите внимание на то, что каждая из них является случайной до тех пор, пока не будет зафиксирован конкретный результат наблюдения.

Случай первый. Объем продаж в следующем квартале характеризуется некоторым числом, значение которого точно неизвестно, но находится среди некоторого набора значений.

Случай второй. Количество устройств с неисправностями, которые будут выпущены на следующей неделе.

Случай третий. Количество квалифицированных специалистов, которые откликнутся на размещенное объявление о найме нового сотрудника.

Случай четвертый. Цена барреля нефти в следующем году.

Случай пятый. Доход следующей семьи, которая ответит на вопрос проводимого исследования.

Случайную величину можно определить как описание численного результата случайного эксперимента. Само значение называют **наблюдаемым значением**. Например, "объем продаж в следующем квартале" — это случайная величина, определяющая и описывающая число, которое будет получено в случайном эксперименте, смысл которого заключается в том, чтобы подождать конца следующего квартала и посчитать объем продаж. Реальное значение, которое может быть получено в будущем, например $3955486, представляет собой наблюдаемое значение такой случайной величины. Обратите внимание на различие между случайной величиной (относящейся к случайному процессу) и наблюдаемым значением (фиксированным числом, которое регистрируется при наблюдении этого процесса). Правило определения вероятностей значений случайной величины называется **распределением вероятностей**.

Многие случайные величины можно охарактеризовать с помощью среднего значения и стандартного отклонения.[1] Кроме того, существует определенная вероятность для каждого из связанных со случайной величиной событий. Мы будем рассматривать случайные величины (случайные переменные) двух типов: *дискретные* и *непрерывные*. С дискретными случайными величинами работать проще, поскольку для них можно составить перечень всех возможных значений. Мы изучим здесь два частных случая распределений, которые оказываются особенно полезными при практическом применении: *биномиальное распределение* (дискретное) и *нормальное распределение* (непрерывное). Более того, часто при вычислении вероятностей можно использовать (что значительно проще) нормальное распределение как достаточно хорошее приближение к распределению биномиальному.

Поскольку существует большое количество разных способов получения данных, существует и много различных типов случайных величин. Примером, не покрывающим даже верхушку этого айсберга, могут служить *экспоненциальное* распределение и распределение *Пуассона*.

Случайная величина называется **дискретной**, если можно перечислить все возможные значения, которые она может получать в результате наблюдений. Случайная величина называется **непрерывной**, если она может принимать любое значение из некоторого интервала (например, произвольное положительное число). Не всегда легко можно определить, относится случайная величина к числу дискретных или к числу непрерывных. Например, объем продаж в следующем квартале может

[1] Все рассматриваемые в этой главе случайные величины имеют среднее значение и стандартное отклонение, однако теоретически могут существовать случайные величины, не имеющие ни среднего, ни стандартного отклонения.

составить $385 298,61, или $385 298,62, или $385 298,63, или любое число не более некоторого достаточно большого значения, например, $4 000 000. В строгом понимании— это дискретная случайная величина (поскольку здесь можно перечислить все возможные значения), однако с практической точки зрения, поскольку различия между соседними значениями в таком перечислении очень малы, с этой случайной величиной можно работать как с непрерывной.

7.1. Дискретные случайные величины

Если для дискретной случайной величины известен список всех возможных значений с их вероятностями (что определяет *распределение вероятности*), то о соответствующем процессе, который порождает случайное и неопределенное значение, известно абсолютно все. Воспользовавшись таким списком, можно вычислить любую представляющую интерес характеристику (например, типичное значение или величину риска) или величину вероятности (для любого события, которое определяется наблюдаемым значением).

Приведем несколько примеров дискретных случайных величин.

1. Количество неисправных устройств, которые будут выпущены на следующей неделе. Список возможных значений имеет вид 0, 1, 2, ...
2. Количество имеющих достаточную квалификацию претендентов, которые откликнутся на объявление о найме на работу. Список возможных значений в этом случае также имеет вид 0, 1, 2, ...
3. Размер бюджета проекта, когда выбор производится из четырех вариантов, предусматривающих финансирование в объеме $26 000, $43 000, $54 000 и $83 000. Перечень возможных значений (в тысячах долларов) состоит из чисел 26, 43, 54 и 83.

Такой список возможных значений, снабженный вероятностями появления этих значений, представляет собой распределение вероятности для дискретной случайной величины. Значениями вероятностей должны быть положительные числа (или 0), сумма которых равна 1. Заданное в таком виде распределение позволяет найти среднее значение, стандартное отклонение и вероятность любого из событий, соответствующих этой случайной величине.

Пример. Доходность при различных сценариях развития экономики

На совещании по оценке перспектив фирмы, проведенном с использованием метода "мозговой атаки", состоялось общее обсуждение вопроса о том, что может произойти в будущем. Было решено упростить ситуацию, рассмотрев наилучший сценарий, худший из возможных сценариев и два промежуточных варианта. После продолжительных дискуссий для каждого их этих четырех сценариев было достигнуто общее согласие относительно того, каким может быть приблизительный доход и какова его вероятность. Обратите внимание на то, что эти значения определяют *распределение вероятности* для случайной величины "доход": существует перечень значений и перечень соответствующих вероятностей. Ниже приведена таблица, где в одном столбце указаны значения случайной величины (в данном случае дохода), а в другом — соответствующие значения вероятностей.

Сценарий развития экономики	Доход, млн дол.	Вероятность
Прекрасный	10	0,20
Хороший	5	0,40
Нормальный	1	0,25
Плохой	–4	0,15

Это распределение вероятности легко использовать для вычисления вероятностей всех событий, связанных с доходом. Вероятность того, что доход составит 10 миллионов долларов, равна, например, 0,20. Вероятность получения дохода в размере 5 миллионов долларов или более можно найти следующим образом: 0,20 + 0,40 = 0,60.

Вычисление среднего и стандартного отклонения

Среднее, или ожидаемое значение, дискретной случайной величины представляет собой некоторое определенное число, характеризующее типичное значение этой величины, подобно тому, как некоторый набор данных характеризуется средним значением.[2] Среднее для случайной величины X обозначают либо малой греческой буквой μ (мю), либо $E(X)$ (читается "ожидаемое значение X"). Формула для вычисления среднего имеет следующий вид.

Среднее или ожидаемое значение дискретной случайной величины X

$$\mu = E(X) = \text{сумма (значение, умноженное на вероятность)} = \sum XP(X).$$

Если бы вероятности всех значений были одинаковыми, мы получили бы простое усреднение всех значений. Вообще говоря, среднее значение случайной величины можно определить как взвешенное среднее всех возможных значений, в котором в качестве весов выступают соответствующие вероятности.

В рассмотренном выше примере средний доход вычисляется следующим образом:

Ожидаемый доход = (10 × 0,20) + (5 × 0,40) + (1 × 0,25) + (–4 × 0,15) = 3,65.

Таким образом, ожидаемый доход составляет 3,65 миллионов долларов. Это значение характеризует все различные возможные результаты (10, 5, 1, –4) одним числом, которое учитывает вероятность (правдоподобие) каждого из них.

Стандартное отклонение дискретной случайной величины приблизительно указывает, насколько реальные значения этой случайной величины могут отличаться от среднего. Во многих случаях в коммерческой деятельности стандартное отклонение характеризует *риск*, показывая, насколько неопределенной является ситуация. Стандартное отклонение обозначается σ, что соответствует использованию нами буквы σ для обозначения стандартного отклонения генеральной совокупности. Формула для вычисления стандартного отклонения имеет такой вид

[2] Среднее значение случайной величины, в отличие от среднего значения некоторого набора данных, нельзя найти с помощью простого вычисления среднего арифметического возможных значений.

Стандартное отклонение дискретной случайной величины X

$$\sigma = \sqrt{\text{Сумма (квадрат отклонения, умноженный на вероятность)}} =$$

$$= \sqrt{\sum (X-\mu)^2 P(X)}\,.$$

Обратите внимание на то, что правильный результат нельзя получить, просто воспользовавшись клавишей Σ калькулятора для суммирования значений одного столбца таблицы, поскольку так не будут правильно использованы значения вероятностей.

В приведенном выше примере стандартное отклонение дохода вычисляется следующим образом:

$$\sigma = \sqrt{\left[(10-3,65)^2\,0,20\right]+\left[(5-3,65)^2\,0,40\right]+\left[(1-3,65)^2\,0,25\right]+\left[(-4-3,65)^2\,0,15\right]} =$$

$$= \sqrt{8,064500+0,729000+1,755625+8,778375} =$$

$$= \sqrt{19,3275} =$$

$$= 4,40.$$

Стандартное отклонение в размере $4 400 000 показывает, что в данном случае присутствует значительный риск. Доход вполне может оказаться на $4 400 000 выше или ниже среднего значения в $3 650 000. В табл. 7.1.1 приведены детали расчетов стандартного отклонения.

Таблица 7.1.1. Вычисление стандартного отклонения дискретной случайной величины

Доход	Вероятность	Отклонение от среднего	Квадрат отклонения	Произведение квадрата отклонения на вероятность
10	0,20	6,35	40,3225	8,064500
5	0,40	1,35	1,8225	0,729000
1	0,25	–2,65	7,0225	1,755625
–4	0,15	–7,65	58,5225	8,778375
				Сумма: 19,3275
				Корень квадратный: 4,40

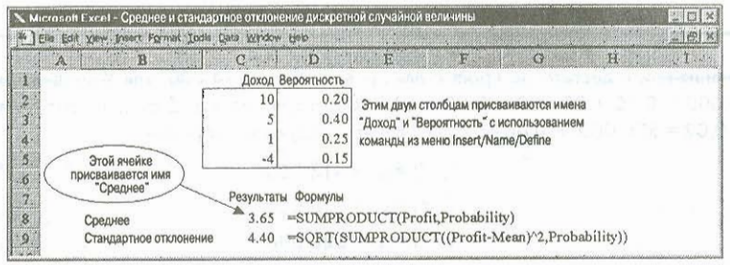
Чтобы вычислить среднее и стандартное отклонения дискретной случайной величины с помощью Excel®, необходимо сделать следующее. Используя команды из меню Excel, дайте имена столбцу значений (в нашем примере он назван Profit (Доход)) и столбцу вероятностей (здесь используется имя Probability (Вероятность)). Среднее (3,65) представляет собой сумму произведений значений на соответствующие вероятности, таким образом, формула для вычисления среднего имеет вид

```
=SUMPRODUCT(Profit, Probability) (= СУММПРОИЗВ(Доход; Вероятность)).
```

Теперь дайте этой ячейке (в ней теперь содержится среднее) имя Mean (Среднее). Стандартное отклонение (4,40) представляет собой квадратный корень (функция Excel SQRT, или КОРЕНЬ) из суммы произведений вероятности на квадрат разности между соответствующим значением и средним. Формула для вычисления стандартного отклонения имеет вид

```
=SQRT(SUMPRODUCT((Profit - Mean)^2, Probability))
=КОРЕНЬ(СУММПРОИЗВ((Доход - Среднее)^2; Вероятность)).
```

Эти формулы дают значение 3,65 для среднего и 4,40 для стандартного отклонения — тот же результат, что и полученный ранее.



На рис. 7.1.1 показано распределение вероятности. Высота вертикальной линии соответствует вероятности, а положение линии — величине дохода для каждого случая. Показано также ожидаемое значение, $3 650 000, и стандартное отклонение, составляющее $4 400 000.



Рис. 7.1.1. Распределение вероятности будущих доходов. Показано также среднее значение (ожидаемый доход) и стандартное отклонение (риск)

Пример. *Оценка риска и доходности*

Перед вами стоит задача — оценить три разных проекта (X, Y и Z) и разработать рекомендации для высшего руководства. По каждому из проектов необходимы инвестиции в объеме $12 000, а возврат средств планируется на следующий год. По проекту X гарантированный возврат составит $14 000. По проекту Y может быть получено либо $10 000, либо $20 000, вероятность в каждом случае составляет 0,5. Проект Z не даст ничего с вероятностью 0,98 или принесет $1 000 000 с вероятностью 0,02. Эти данные собраны в табл. 7.1.2.

Таблица 7.1.2. Возврат средств по трем проектам и соответствующие вероятности

Проект	Возврат	Вероятность
X	14 000	1,00
Y	10 000	0,50
	20 000	0,50
Z	0	0,98
	1 000 000	0,02

Средние значения найти достаточно просто: для проекта X это $14 000, для Y среднее значение находится как 10 000 × 0,50 + 20 000 × 0,50 = $15 000, а для проекта Z среднее составляет 0 × 0,98 + 1 000 000 × 0,02 = $20 000. Мы можем записать это следующим образом:

$$E(X) = \mu_x = \$14\,000$$
$$E(Y) = \mu_y = \$15\,000$$
$$E(Z) = \mu_z = \$20\,000$$

Если исходить только из этих величин, проект Z может показаться самым лучшим, а проект X — худшим из всех. Однако средние значения не дают полной информации. Так, например, несмотря на то, что по проекту Z ожидаемый возврат оказывается самым большим, этот проект несет еще и максимальный риск: вероятность отсутствия каких-либо выплат составляет 98%! Присущие каждому из рассматриваемых случаев риски характеризуются стандартным отклонением.

$$\sigma_x = \sqrt{(14000-14000)^2 \times 1{,}00} = \$0,$$

$$\sigma_y = \sqrt{(10000-15000)^2 \times 0{,}50 + (20000-15000)^2 \times 0{,}50} = \$5000,$$

$$\sigma_z = \sqrt{(0-20000)^2 \times 0{,}98 + (1000000-20000)^2 \times 0{,}02} = \$140000.$$

Исследование стандартного отклонения подтверждает возникшие опасения. Проект Z действительно оказывается самым рискованным — намного более рискованным, чем два других. Проект X — самый безопасный. Это верное дело, не несущее никакого риска. В случае проекта Y риск составляет $5000.

Какой проект выбрать? На этот вопрос нельзя ответить, используя только методы статистического исследования. Несмотря на то что ожидаемое значение и стандартное отклонение предоставляют нам полезные для принятия решения данные, задача этим не исчерпывается. Обычно предпочтение отдается большим ожидаемым выплатам и меньшему риску. Однако в приведенном примере возможность получения больших выплат сопряжена с более высоким риском. Окончательный выбор проекта требует от вас (и от вашей фирмы) определить, что важнее — доход или риск, — и исходя их этого определить, оправдывает увеличение ожидаемых выплат такое увеличение риска или нет.[3]

Что если попробовать оценить проекты в терминах дохода, а не выплат? Поскольку каждый проект требует инвестиций в $12000, для перехода от выплат к доходу необходимо вычесть $12000 из каждой величины выплат в таблице распределения вероятностей.

$$\text{Доход} = \text{Выплаты} - \$12\,000.$$

Применяя правила из раздела 5.2, которые так же справедливы для характеристик случайных величин, как и для характеристик наборов данных, вычтем $12 000 из каждого среднего значения и оставим без изменений стандартные отклонения. Таким образом, без дополнительных подробных вычислений мы получаем следующие величины ожидаемого дохода:

X: $2 000
Y: $3 000
Z: $8 000

Стандартные отклонения для дохода такие же, как и для размера выплат:

X: $0
Y: $5 000
Z: $140 000

[3] В курсе теории финансов вы, видимо, встречались с другим часто используемым в оценке проектов фактором — корреляцией (если таковая есть) между случайными величинами выплат по проекту и доходностью рыночного портфеля. Это позволяет измерить диверсифицируемый и недиверсифицируемый риск проекта. Корреляция (статистическая мера связи) будет обсуждаться в главе 11.

7.2. Биномиальное распределение

Проценты очень часто используют в коммерческих расчетах. Если количество наступлений события выражается как процент от общего количества возможностей, то количество наступлений события должно иметь *биномиальное распределение*. В этом случае существует ряд сохраняющих время и силы способов вычисления ожидаемого значения, стандартного отклонения, а также вероятностей различных событий. Иногда интерес представляет сам процент; в других случаях более полезным может быть количество наступлений события. Биномиальное распределение в любом случае поможет найти требуемую величину. Ниже приведены примеры некоторых случайных величин, имеющих биномиальное распределение.

1. Количество заказов, которые будут получены в результате следующих трех телефонных звонков в отдел торговли по каталогам.
2. Количество имеющих дефекты изделий среди 10 единиц выпущенной продукции.
3. Количество среди 200 опрошенных людей, выразивших желание покупать данный товар.
4. Количество акций, курс которых вчера повысился, среди всех тех акций, торговля которыми ведется на основных биржах.
5. Количество женщин, работающих в отделе, в штате которого насчитывается 75 человек.
6. Количество голосов, которые будут отданы за республиканцев (или демократов) на следующих выборах.

Определение биномиального распределения и биномиального соотношения

Рассмотрим некоторое конкретное событие. Каждый раз при выполнении случайного эксперимента это событие или происходит, или нет. Наличие двух возможных результатов и определяет приставку *би-* (т.е. "двух-") в слове *биномиальное*. Случайная величина X, которая представляет собой *число наступлений* определенного события в результате n попыток, имеет **биномиальное распределение** в следующих случаях.

1. Если в каждой из n попыток вероятность наступления события π одна и та же.
2. Если все попытки независимы друг от друга.

Требование независимости означает невозможность "заглядывать", как, например, в случае распределения заказов особого блюда посетителями ресторана. Если некоторые из посетителей делают заказ этого блюда потому, что видят, как другие явно им наслаждаются (мысленно восклицая: "Я тоже хочу это!"), количество заказов *не* будет соответствовать биномиальному распределению. Для того чтобы распределение заказов оказалось биномиальным, все посетители должны делать свой выбор независимо.

Биномиальная пропорция p — это представление имеющей биномиальное распределение случайной переменной X в виде доли общего количества попыток n.

Биномиальная пропорция

$$p = \frac{X}{n} = \frac{\text{Количество наступлений события}}{\text{Количество попыток}}.$$

(Обратите внимание на то, что π — это фиксированное число, определяющее вероятность наступления события, а p — случайная величина, зависящая от наблюдаемых данных.) Например, если в результате опроса n = 600 покупателей было установлено, что X = 38 из них собираются покупать ваш товар, биномиальная пропорция оказывается равной

$$p = \frac{X}{n} = \frac{38}{600} = 0,063, \text{ или } 6,3\%.$$

Выражающее биномиальную пропорцию (соотношение) число p называют также *биномиальной долей*. Его можно воспринимать еще и как относительную частоту, с которой мы уже встречались в главе 6.

Пример. *Сколько подано заявок? Сложный путь вычислений*

В этом примере показан сложный способ исследования биномиальной случайной величины. Несмотря на то что необходимость в построении дерева вероятностей возникает достаточно редко, поскольку оно обычно довольно велико, на него полезно один раз посмотреть, чтобы понять, что же в действительности происходит в случае биномиального распределения. Более того, когда будет представлен более короткий (более легкий) способ вычислений, вы поймете с благодарностью, как много времени вы экономите!

Предположим, что нас интересуют следующие n = 3 звонка в отдел торговли по каталогам, а из опыта известно (или мы так предполагаем[4]), что π = 0,6, т.е. 60% обращений приводят к получению заказа на покупку (другие звонки связаны с запросом информации или неправильным соединением). Что можно сказать о количестве обращений, которые приведут в данном случае к оформлению заказа? Понятно, что это число окажется равным 0, 1, 2 или 3. Поскольку каждый звонок скорее приведет к подаче заявки, чем нет, следует, видимо, ожидать, что вероятность подачи трех заявок будет больше, чем вероятность того, что не будет подано ни одной заявки. Но как найти эти вероятности? Дерево вероятностей на рис. 7.2.1а дает полное описание ситуации и иллюстрирует результат каждого из трех телефонных звонков.

[4] При решении задач с биномиальным распределением вероятность π обычно дается в условии. В реальной жизни у этой вероятности те же источники, о которых мы вели речь ранее: относительная частота, теоретические расчеты и субъективные оценки.

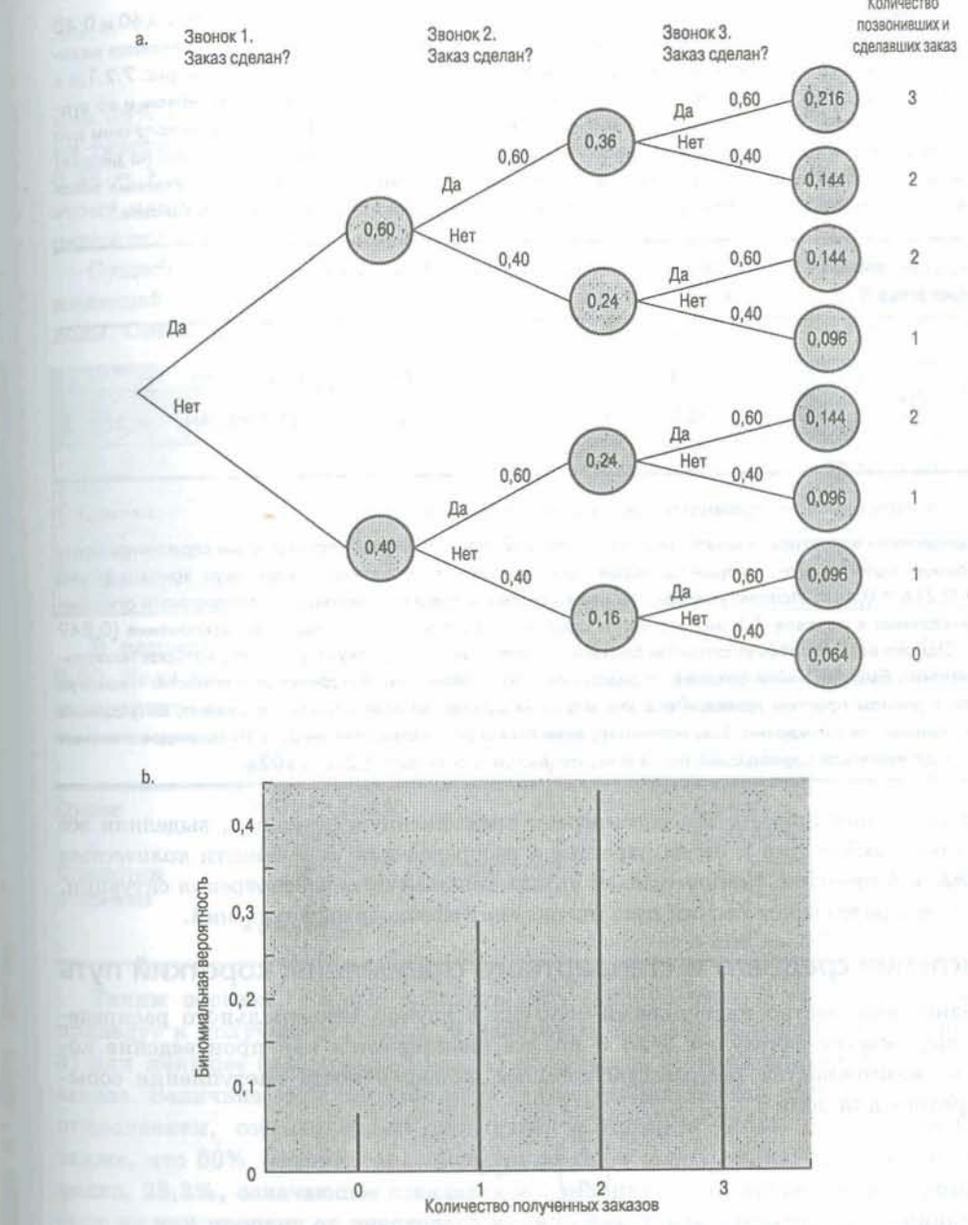


Рис. 7.2.1. а) Дерево вероятностей для последовательных телефонных звонков, каждый из которых либо приводит, либо не приводит к получению заказа. Существует восемь возможных комбинаций (кружки в крайнем правом столбце). В частности, видно, что имеются три способа получения двух заказов: второй, третий и пятый сверху кружки, что дает вероятность 3 × 0,144 = 0,432; б) Биномиальное распределение вероятности количества звонков, приводящих к получению заказа

Обратите внимание на то, что показанные вдоль ветвей условные вероятности всегда равны 0,60 и 0,40 (это вероятности для каждого из телефонных звонков), поскольку мы считаем, что заказы делаются независимо и звонящие в отдел лица не влияют друг на друга. Количество заказов показано на рис. 7.2.1,а в крайнем правом столбце; например, второе сверху число, 2, отражает тот факт, что в первом и во втором телефонном звонках (но не в третьем) были сделаны заказы и, таким образом, были получены два заказа. Обратите внимание на то, что существуют три способа (и, соответственно, три пути на дереве) получить два заказа. Для того чтобы построить распределение вероятности количества полученных заказов, можно сложить все вероятности различных способов реализации данного количества заказов.

Число позвонивших и сделавших заказ X	Процент тех, кто сделал заказ $p = X/n$	Вероятность
0	0,0	0,064
1	33,3	0,288 (=0,096 + 0,096 + 0,096)
2	66,7	0,432 (=0,144 + 0,144 + 0,144)
3	100,0	0,216

Полученное распределение вероятности показано на рис. 7.2.1,б.

Из распределения вероятности можно найти любую необходимую вероятность простым сложением соответствующих значений. Так, например, вероятность получения по меньшей мере двух заказов равна 0,432 + 0,216 = 0,648. Используя формулы для вычисления среднего значения и стандартного отклонения, приведенные в разделе 7.1, можно найти среднее (1,80 заказа) и стандартное отклонение (0,849 заказа). Однако все это требует слишком больших затрат труда. Существует формула, которая позволяет значительно быстрее найти среднее, стандартное отклонение и необходимые вероятности. Несмотря на то что в данном простом примере все эти значения можно вычислить непосредственно, ситуация не всегда складывается так удачно. Так, например, если бы мы рассматривали не 3, а 10 последовательных звонков, в правой части дерева было бы не 8 вероятностей, как на рис. 7.2.1, а 1024.

В приведенном примере мы рассмотрели предложенную ситуацию, выделили все возможные комбинации и затем перешли к распределению вероятности *количества наступлений события*. Концептуально это правильный путь рассмотрения ситуации. А теперь мы рассмотрим легкий путь вычислений необходимых значений.

Вычисление среднего и стандартного отклонения: короткий путь

Среднее количество наступлений события в случае биномиального распределения выражается формулой E(X) = nπ, т.е. вычисляется как произведение количества возможностей реализации события на вероятность наступления события. Среднее для доли

$$E\left(\frac{X}{n}\right) = E(p) = \pi$$

равно вероятности наступления события.[5]

Этого и следовало ожидать. Например, если опрошена выборка из 200 избирателей и для каждого из них вероятность отдать предпочтение вашему кандидату равна 58%, то в среднем следует ожидать, что

[5] Можно воспринимать X/n как относительную частоту события. Тот факт, что $E(X/n)$ равно π, свидетельствует, что усредненная относительная частота события равна вероятности этого события.

$$E\left(\frac{X}{n}\right) = E(p) = \pi = 0,58,$$

или 58% опрошенных избирателей отдадут предпочтение вашему кандидату. Если говорить о количестве опрошенных, следует ожидать, что $E(X) = n\pi = 200 \times 0,58 = 116$ человек из 200 попавших в данную выборку отдадут предпочтение этому кандидату. Конечно, реальное количество и процент будут случайным образом отличаться от этих ожидаемых значений.

Существуют также формулы для вычисления стандартного отклонения имеющей биномиальное распределение величины и для вычисления процентной доли. Они приведены с формулой для вычисления ожидаемого значения ниже.

Среднее значение и стандартное отклонение для биномиального распределения

	Количество наступлений события, X	Доля или процент, p = X/n
Среднее	$E(X) = \mu_X = n\pi$	$E(p) = \mu_p = \pi$
Стандартное отклонение	$\sigma_X = \sqrt{n\pi(1-\pi)}$	$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

В рассмотренном выше примере о приеме заказов по телефону n =3 и π = 0,60. Если воспользоваться приведенными выше формулами, среднее значение и стандартное отклонение можно найти следующим образом.

	Количество наступлений события, X	Доля или процент, $p = X/n$
Среднее	$E(X) = n\pi = 3 \times 0,60$ = 1,80 звонков	$E(p) = \pi$ = 0,60, или 60%
Стандартное отклонение	$\sigma_X = \sqrt{n\pi(1-\pi)} = \sqrt{3 \times 0,60(1-0,60)}$ = 0,849 звонков	$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0,60(1-0,60)}{3}}$ = 0,283, или 28,3%

Таким образом, следует ожидать, что 1,80 из этих 3 телефонных звонков приведут к получению заказов. В некоторых случаях большее (например, 2 или 3) или меньшее (например, 0 или 1) количество звонков приведут к получению заказа. Величина этой неопределенности определяется, как обычно, стандартным отклонением, составляющим для данного примера 0,849 звонка. Ожидается также, что 60% из этих трех звонков приведут к получению заказов. Последнее число, 28,3%, означающее стандартное отклонение для процента, интерпретируется не как процент от некоторого количества, а как *процентные единицы*. Это означает, что если мы ожидаем 60%, то реально мы можем наблюдать на 28,8 процентных единиц больше (т.е. 60 + 28,3 = 88,3%) или ниже (т.е. 60 – 28,3 = 31,7%) этого значения. Это естественно, поскольку, как известно, стандартное отклонение выражается в тех же единицах, что и данные, а в случае процентной доли p стандартное отклонение выражается именно в процентных единицах (т.е. собственно в процентах).

Пример. *Запоминание рекламы*

Предположим, что ваша организация собирается заключить контракт с фирмой, специализирующейся на маркетинговых исследованиях, для изучения воздействия вашей рекламы на жителей Америки. В определенный день отобранные люди должны прийти и просмотреть телепрограммы с рекламой (многих товаров различных компаний). На следующий день эти люди придут еще раз, чтобы ответить на ряд вопросов. В частности, планируется определять уровень *запоминания*, который вычисляется как процент людей, которые вспомнят рекламу вашей фирмы на следующий день после просмотра.

До заключения контракта на проведение этой работы необходимо оценить, насколько достоверными и точными будут полученные результаты. Средства, выделяемые вашей фирмой на исследование рекламы, дают возможность пригласить для участия в исследовании 50 человек. В результате обсуждения с представителями фирмы стало известно, что есть смысл предположить, что 35% участников будут помнить рекламу; однако точная доля, естественно, неизвестна. Насколько точными будут результаты исследования, если исходить из предположения, что доля таких людей действительно составит 35%? Это означает: насколько будет отличаться полученный в результате опроса процент людей, запомнивших рекламу, от предполагаемого значения $\pi = 0{,}35$ при $n = 50$ и биномиальном распределении результатов? Ответ находим, вычисляя соответствующее стандартное отклонение:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0{,}35(1-0{,}35)}{50}} = 0{,}0675, \text{ или } 6{,}75\%.$$

Это означает, что полученный при проверке запоминания результат (процент участвующих в исследовании людей, которые запомнят рекламу) будет, скорее всего, отличаться от истинного значения процентной доли для всего населения примерно на 7% в любую (в большую или в меньшую) сторону.

Однако ваша фирма считает, что необходимо получить более точные результаты. Повысить точность результатов можно за счет сбора большего объема информации, для чего необходимо увеличить размер выборки n. В результате проверки бюджета исследования и обсуждения затрат установлено, что можно привлечь к исследованию $n = 150$ человек. Для такой большей выборки стандартное отклонение уменьшится:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0{,}35(1-0{,}35)}{150}} = 0{,}0389, \text{ или } 3{,}89\%.$$

Кажется странным, что увеличение затрат не привело к значительному улучшению результата. При увеличении размера исследования в три раза точность не возросла даже вдвое! Это связано с тем, что в формулу входит не сама величина n, а *квадратный корень из* n. Однако, несмотря на это, фирма решает пойти на дополнительные затраты, считая, что достигнутая дополнительная точность стоит того.

Вычисление вероятностей

Рассмотрим биномиальное распределение, для которого известны величины n и π; и необходимо найти вероятность того, что X будет в точности равно некоторому значению a. Существует формула для вычисления вероятности, которая оказывается полезной в случае малых и средних значений n. (При больших n можно использовать более простое, по сравнению с представленным здесь строгим решением, основанное на нормальном распределении приближение, которое мы рассмотрим в разделе 7.4.) Кроме того, в табл. В.3 приложения В приводятся точные значения вероятностей биномиального распределения и кумулятивные вероятности для n от 1 до 20 и $\pi = 0{,}05$, 0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9 и 0,95. Точная формула имеет следующий вид.[6]

Вероятность того, что при биномиальном распределении X равно a

$$P(X=a) = \binom{n}{a}\pi^a(1-\pi)^{n-a} = \frac{n!}{a!(n-a)!}\pi^a(1-\pi)^{n-a} = \frac{1\times2\times3\times\ldots\times n}{[1\times2\times3\times\ldots\times a][1\times2\times3\times\ldots\times(n-a)]}\pi^a(1-\pi)^{n-a}.$$

Если применить эту формулу для каждого из значений a от 0 до n (иногда этот процесс оказывается очень трудоемким), можно полностью вычислить распределение вероятности. Затем, складывая необходимые из полученных таким образом значений, можно вычислить вероятность для любого интересующего нас значения X.

Рассмотрим пример применения приведенной выше формулы. Предположим, что существуют $n = 5$ возможных попыток, причем вероятность успеха каждой из них равна $\pi = 0{,}8$. Найдем вероятность в точности $a = 3$ успешных попыток. Ответ вычисляем следующим образом:

$$P(X=3) = \binom{5}{3}0{,}8^3(1-0{,}8)^{5-3} = \frac{5!}{3!(5-3)!}0{,}8^3\times0{,}2^2 = \frac{1\times2\times3\times4\times5}{(1\times2\times3)(1\times2)}0{,}512\times0{,}040 = 10\times0{,}02048 = 0{,}2048.$$

[6] Обозначение $n!$ читается как "n факториал" и означает произведение целых чисел от 1 до n. Например, $4! = 1\times2\times3\times4 = 24$. (Принято также считать, что $0! = 1$.) Многие калькуляторы имеют отдельную кнопку для вычисления факториалов, которая работает для чисел от 0 до 69. Выражение

$$\binom{n}{a} = \frac{n!}{a!(n-a)!}$$

используется для обозначения биномиального коэффициента, читается как "выбор из n по a" и представляет количество различных способов выбора a результатов из n возможных. Например, при $n = 5$ и $a = 3$ биномиальный коэффициент равен

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{120}{6\times2} = 10.$$

Например, существует 10 способов (комбинаций) выбрать из пяти человек трое таких, которые готовы купить некоторый товар: это могут быть три первых человека, два первых и четвертый и т. д.

Это вероятность *в точности* трех успешных попыток. Если необходимо найти также вероятность достижения успеха *три или более раз*, необходимо провести расчеты по этой формуле еще два раза: один раз при a = 4 и один раз при a =5; вероятность достижения успеха три или более раз будет равна сумме полученных значений. Для вычисления вероятностей можно также воспользоваться компьютером, как показано ниже.

Функция плотности вероятности и функция кумулятивного распределения
Биномиальное распределение с n =5 и p = 0,800000

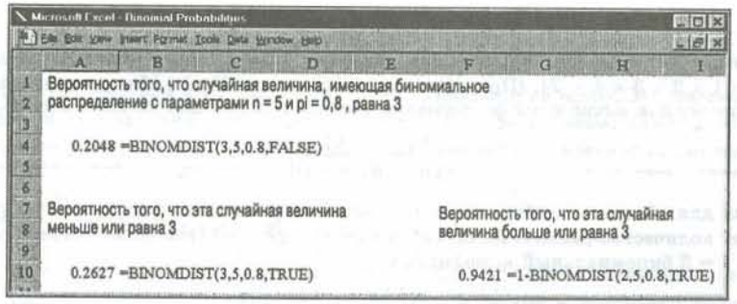
a	P(X=a)	P(X<=a)
0	0,0003	0,0003
1	0,0064	0,0067
2	0,0512	0,0579
3	0,2048	0,2627
4	0,4096	0,6723
5	0,3277	1,0000

После вычисления отдельных значений вероятности (для 3, 4 и 5 успешных попыток) получаем искомый результат:

$$P(X \geq 3) = P(X=3) + P(X=4) + P(X=5) = 0,2048 + 0,4096 + 0,3277 = 0,9421.$$

Таким образом, вероятность достижения успеха три или более раз из пяти возможностей составляет 94,2%. Кроме того, в соответствии с правилом дополнительности, вероятность *три или более* должна быть равна единице минус вероятность *два или менее*, которая представлена среди результатов компьютерных вычислений числом 0,0579. Используя это значение, находим ответ: 1 – 0,0579 = 0,9421.

При вычислении вероятностей биномиального распределения с помощью Excel, чтобы получить вероятность P(X=a) того, что результат окажется равным *a*, необходимо использовать формулу =BINOMDIST(a, n, π, FALSE) (=БИНОМРАСП(a; n; π; ЛОЖЬ)), а для вычисления вероятности P(X≤a) того, что результат окажется *меньше чем, или равным* a, использовать формулу =BINOMDIST(a, n, π, TRUE)[7] (=БИНОМРАСП(a; n; π; ИСТИНА)). Вычисления для рассмотренного примера приведены ниже.



[7] Значения FALSE (ложь) и TRUE (истина) в формулах обработки биномиальных распределений в Excel определяют, является ли распределение вероятностей кумулятивным.

Пример. Сколько крупных клиентов позвонят завтра?

Сколько крупных клиентов вашей фирмы (всего у вас есть 6 крупных клиентов) позвонят завтра? Вы склонны предполагать, что каждый из них может позвонить с вероятностью π = 0,25 и что обращаются к вам они независимо друг от друга. Если так, то количество крупных клиентов, которые позвонят завтра, имеет биномиальное распределение.

Сколько звонков от крупных клиентов можно ожидать? Другими словами, чему равно ожидаемое значение *X*? Вот ответ на этот вопрос: $E(X) = n \times \pi = 1,5$ звонков от крупных клиентов. Стандартное отклонение составляет $\sigma_x = \sqrt{6 \times 0,25 \times (1-0,25)} = 1,060660$. Это означает, что следует ожидать на один или два звонка больше или меньше значения 1,5, полученного в результате вычислений. Несмотря на то что это значение дает некоторое представление о количестве звонков, которое можно ожидать, полученный результат еще не характеризует вероятность того, что такое количество клиентов действительно позвонят. Вычислим теперь соответствующие вероятности.

Чему равна вероятность того, что позвонят именно двое из шести крупных клиентов? Ответ на этот вопрос имеет следующий вид:

$$P(X=2) = \binom{6}{2} 0,25^2 (1-0,25)^{6-2} = 15 \times 0,0625 \times 0,316406 = 0,297.$$

Ниже приведено полное распределение вероятности для количества звонков от крупных клиентов, которые поступят завтра, включая все возможные случаи от 0 до n =6.

Функция плотности вероятности и функция кумулятивного распределения
Биномиальное распределение с n =6 и p = 0,250000

a	P(X=a)	P(X<=a)
0	0,1780	0,1780
1	0,3560	0,5339
2	0,2966	0,8306
3	0,1318	0,9624
4	0,0330	0,9954
5	0,0044	0,9998
6	0,0002	1,0000

Обратите внимание, что наиболее вероятными оказываются один или два звонка, как и следовало ожидать при среднем значении 1,5 звонка.

Воспользовавшись этим распределением вероятности, можно вычислить любую представляющую интерес вероятность, характеризующую обращение крупных клиентов в фирму завтра по телефону. Вероятность того, что позвонят все шесть клиентов, невелика (0,0002, или 0,02%, это значительно меньше вероятности в 1%). Вероятность того, что позвонят 4 или более из них, составляет 0,0330 + 0,0044 + 0,0002 = 0,0376. Вероятность того, что день будет совершенно спокойный, без звонков от крупных клиентов, составляет 0,178. Полное распределение вероятности для данного случая показано на рис. 7.2.2.

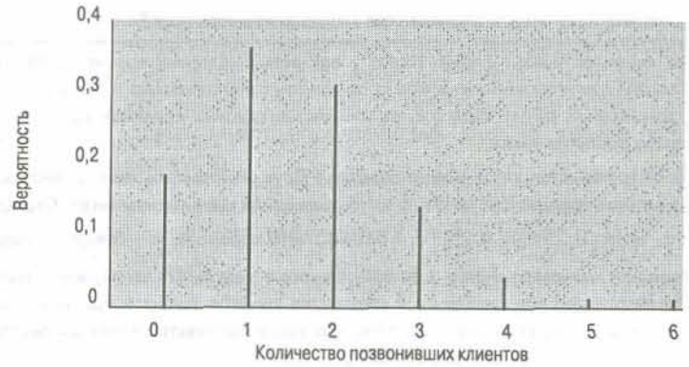


Рис. 7.2.2. Распределение вероятности для количества крупных клиентов, которые позвонят завтра. Высота вертикальных столбцов соответствует вероятностям для биномиального распределения, вычисленным с использованием соответствующей формулы при n = 6 и π = 0,25. Величина a отложена вдоль горизонтальной оси

Пример. Сколько логических анализаторов запланировать для производства?

Качеству выпускаемого товара уделяется большое внимание, однако логические анализаторы настолько сложны, что их производство все еще не лишено определенных недостатков. Из предыдущего опыта известно, что около 97% готовой продукции работает хорошо. Сегодня предстоит отгрузить потребителю 17 таких устройств. Вопрос состоит в следующем: сколько устройств надо выпустить, чтобы иметь разумные гарантии того, что можно будет отгрузить 17 работающих должным образом логических анализаторов?

Разумно предположить, что количество работающих устройств подчиняется биномиальному распределению, где n — количество запланированных для производства устройств, а π- вероятность того, что каждое из них работает должным образом (эта вероятность равна 0,97). Такое предположение даст возможность рассчитать вероятность надлежащего функционирования 17 или более устройств, выпущенных в соответствии с производственным планом.

Что произойдет, если в плане предусмотреть выпуск 17 устройств, не предусмотрев возможности дефектов? В этом случае может показаться, что высокий уровень выхода годного товара (97%) сыграет свою положительную роль, но в действительности вероятность того, что все 17 устройств будут работать хорошо (при n =17 и a =17), составит только 0,596:

$$P(X = 17 \text{ работающих устройств}) = \binom{17}{17} 0{,}97^{17} 0{,}03^{0} =$$

$$= 1 \times 0{,}595826 \times 1 =$$

$$= 0{,}596.$$

Таким образом, если в производственный план внести то же количество товара, что и необходимо отгрузить, 17 единиц, риск окажется очень большим! Вероятность того, что удастся выполнить заказ, окажется равной только 59,6%, а с вероятностью 40,4% отправить товар соответственно заказу и в рабочем состоянии не удастся. Распределение вероятности для данного случая показано на рис. 7.2.3.

Похоже, что лучше включить в план выпуск не 17, а большего количества устройств. Что если в производственный план внести выпуск n = 18 устройств? Для того чтобы найти вероятность того, что удастся отгрузить *по меньшей мере* 17 работающих должным образом устройств, необходимо сложить вероятности для a = 17 и a =18:

$$P(X \geq 17) = P(X = 17) + P(X = 18) =$$

$$= \binom{18}{17} 0{,}97^{18} 0{,}03^{1} + \binom{18}{18} 0{,}97^{18} 0{,}03^{0} =$$

$$= 18 \times 0{,}595826 \times 0{,}03 + 1 \times 0{,}577951 \times 1 =$$

$$= 0{,}322 + 0{,}578 =$$

$$= 0{,}900.$$

Таким образом, если в производственный план заложить выпуск 18 устройств, вероятность успешной отгрузки 17 работающих устройств составит 90%. Этот результат выглядит неплохо, однако все еще необходимо учитывать то, что возможность неудачи составляет 10%. Соответствующее распределение вероятности показано на рис. 7.2.4.

Аналогичные расчеты показывают, что если внести в производственный план выпуск 19 устройств, то вероятность получить для отгрузки потребителю 17 работающих устройств составит 98,2% (9,2% + 32,9% + 56,1%). Таким образом, для того, чтобы в достаточной мере гарантировать успешное выполнение заказа на 17 работающих устройств, в план следует внести выпуск по меньшей мере 19 устройств.

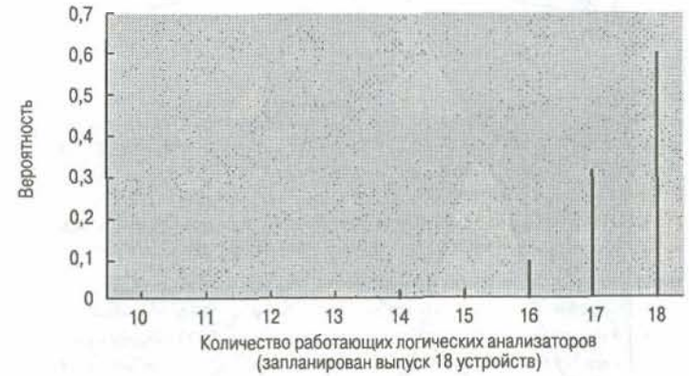


Рис. 7.2.4. Распределение вероятности для количества работающих логических анализаторов, если запланирован выпуск 18 устройств. Это биномиальное распределение с n =18 и π = 0,97

7.3. Нормальное распределение

Из главы 3 вы уже узнаете, как можно определить, имеет ли некоторый набор данных приблизительно нормальное распределение. Теперь мы выясним, как вычислять вероятности для этого уже знакомого нам колоколообразного распределения. Одна из причин практической полезности использования нормального распределения состоит в том, что, зная только среднее значение и стандартное отклонение, можно вычислить любую представляющую интерес вероятность (конечно, при условии, что распределение — действительно нормальное).

Нормальное распределение — это непрерывное распределение, имеющее графическое представление в виде хорошо известной симметричной колоколообраз-

ной кривой, показанной на рис. 7.3.1,а. Для любой комбинации значений среднего и положительного стандартного отклонения можно начертить свой график.[8] Кривая сдвигается вправо или влево так, чтобы вершина "колокола" располагалась над средним значением, и растягивается или сжимается так, чтобы масштаб по горизонтали соответствовал стандартному отклонению. На рис. 7.3.1,б показаны два различных нормальных распределения.

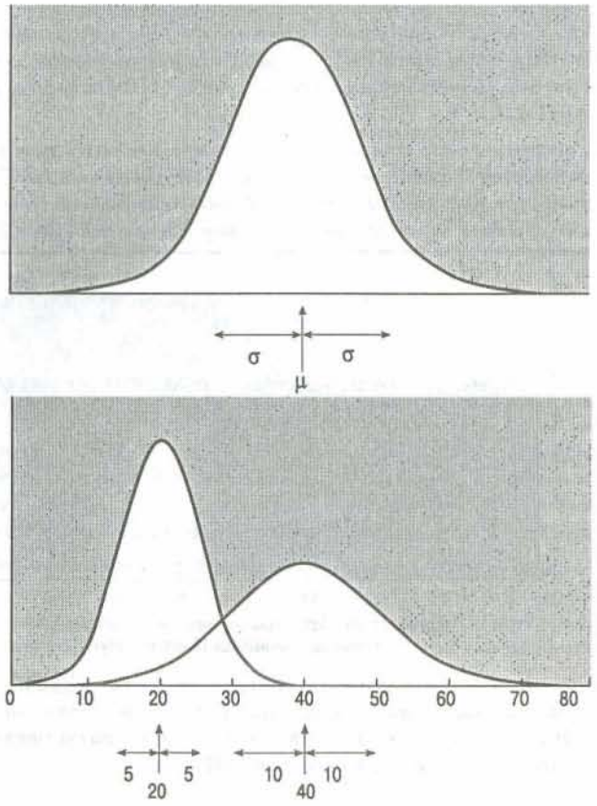


Рис. 7.3.1. а) Нормальное распределение со средним значением μ и стандартным отклонением σ. Обратите внимание на то, что среднее может быть любым числом, а стандартное отклонение — любым положительным числом; б) Два различных нормальных распределения. Кривой, которая располагается левее, соответствует меньшее среднее значение (20) и меньшее стандартное отклонение (5). Расположенной правее более пологой кривой соответствует среднее значение 40 и стандартное отклонение 10

[8] Формула, описывающая нормальное распределение вероятности со средним значением μ и стандартным отклонением σ, имеет вид $\frac{1}{\sqrt{2\pi}\sigma} e^{-[(x-\mu)/\sigma]^2/2}$.

Представление вероятности как площади под кривой

Колоколообразная кривая позволяет наглядно представить вероятность для случая нормального распределения. Более правдоподобным оказывается наблюдение значений, расположенных вблизи центра кривой, там, где она подымается выше. Вблизи краев, где кривая проходит ниже, наблюдение соответствующих значений оказывается менее правдоподобным. Строго говоря, вероятность того, что значение попадет в некоторый интервал на числовой прямой, равна *площади соответствующей области под кривой*, как показано на рис. 7.3.2.

Обратите внимание на то, что заштрихованная область, расположенная вблизи центра кривой, имеет большую площадь, чем область такой же ширины, но расположенная ближе к краю. Для того чтобы увидеть это, сравните рис. 7.3.2 и 7.3.3.

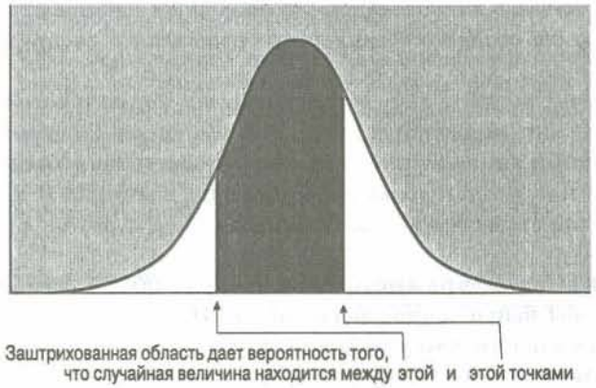


Рис. 7.3.2. Вероятность того, что имеющая нормальное распределение случайная величина принимает значения, лежащие в некотором интервале, равна площади под кривой нормального распределения между значениями, ограничивающими этот интервал. Вероятность наблюдения тех значений, которые ближе расположены к среднему, оказывается выше

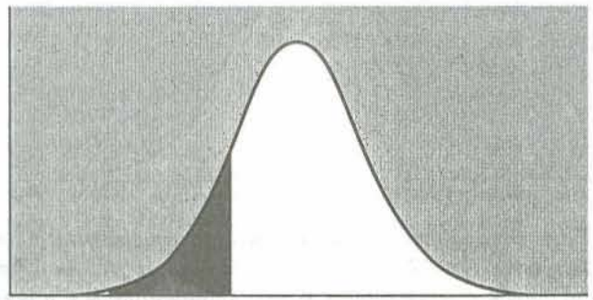


Рис. 7.3.3. Вероятность попадания в интервал, расположенный дальше от центра кривой. Поскольку здесь кривая нормального распределения проходит ниже, вероятность оказывается меньше, чем представленная на рис. 7.3.2

Стандартное нормальное распределение *Z* и соответствующие вероятности

Стандартное нормальное распределение — это нормальное распределение со средним значением $\mu = 0$ и стандартным отклонением $\sigma = 1$. Для обозначения случайной величины, имеющей стандартное нормальное распределение, часто используют букву *Z*. Один из способов вычисления вероятностей для нормального распределения состоит в использовании таблиц вероятностей для стандартного распределения, поскольку совершенно нереально заготовить таблицы для каждой возможной комбинации среднего и стандартного отклонения. С помощью стандартного нормального распределения можно представить любое нормальное распределение, если рассмотрение вести не в реальных единицах измерения (например, в долларах), а в *количествах стандартных отклонений в большую или меньшую сторону от среднего*. Стандартное нормальное распределение показано на рис. 7.3.4.

Таблица вероятностей для стандартного нормального распределения (табл. 7.3.1) содержит вероятности того, что имеющая стандартное нормальное распределение случайная величина *Z* принимает значение *меньше* некоторого заданного числа *z*. Например, вероятность того, что величина *Z* меньше 1,38, равна 0,9162; это проиллюстрировано площадью под кривой распределения на рис. 7.3.5.

Действительно, видно, что заштриховано около 90% всей площади под кривой. Для того чтобы найти точное значение (0,9162), в таблице вероятностей для стандартного нормального распределения необходимо найти z = 1,38 и посмотреть, какая вероятность соответствует этому значению. Найдите также строки для значений 2,35 (соответствующая вероятность 0,9906), 0 (соответствующая вероятность 0,5000) и −082 (соответствующая вероятность 0,2161). А какая вероятность соответствует значению z = 0,36?

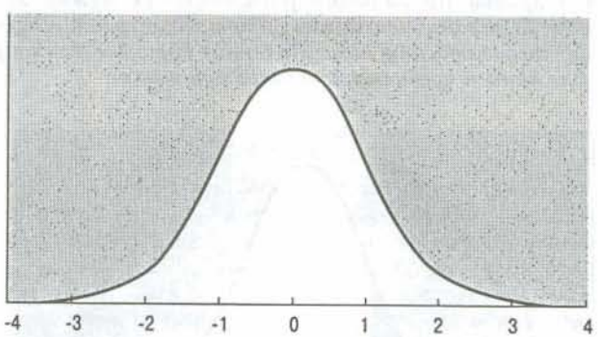


Рис. 7.3.4. Стандартное нормальное распределение случайной величины Z со средним значением $\mu = 0$ и стандартным отклонением $\sigma = 1$. Стандартным нормальным распределением можно пользоваться для исследования любого нормального распределения, если вести речь в терминах количества стандартных отклонений от среднего в большую или меньшую сторону

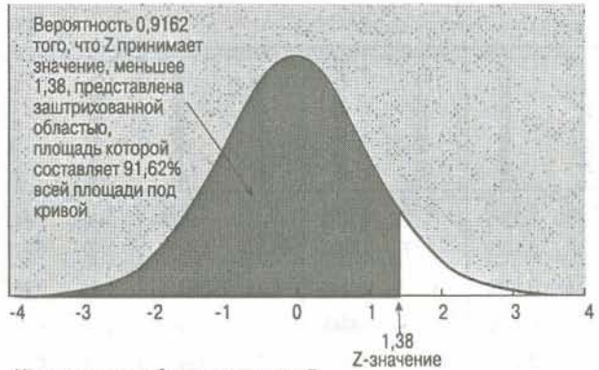


Рис. 7.3.5. Вероятность того, что имеющая стандартное нормальное распределение случайная величина меньше, чем *z = 1,38, равна 0,9162. Это видно из таблицы вероятностей для стандартного нормального распределения. Данная вероятность соответствует области, заштрихованной слева от числа 1,38, площадь которой составляет 91,62% всей площади под кривой*

Таблица 7.3.1. Таблица вероятностей для стандартного нормального распределения

Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность
−2,00	0,0228	−1,00	0,1587	−0,00	0,5000	0,00	0,5000	1,00	0,8413	2,00	0,9772
−2,01	0,0222	−1,01	0,1562	−0,01	0,4960	0,01	0,5040	1,01	0,8438	2,01	0,9778
−2,02	0,0217	−1,02	0,1539	−0,02	0,4920	0,02	0,5080	1,02	0,8461	2,02	0,9783
−2,03	0,0212	−1,03	0,1515	−0,03	0,4880	0,03	0,5120	1,03	0,8485	2,03	0,9788
−2,04	0,0207	−1,04	0,1492	−0,04	0,4840	0,04	0,5160	1,04	0,8508	2,04	0,9793
−2,05	0,0202	−1,05	0,1469	−0,05	0,4801	0,05	0,5199	1,05	0,8531	2,05	0,9798
−2,06	0,0197	−1,06	0,1446	−0,06	0,4761	0,06	0,5239	1,06	0,8554	2,06	0,9803
−2,07	0,0192	−1,07	0,1423	−0,07	0,4721	0,07	0,5279	1,07	0,8577	2,07	0,9808
−2,08	0,0188	−1,08	0,1401	−0,08	0,4681	0,08	0,5319	1,08	0,8599	2,08	0,9812
−2,09	0,0183	−1,09	0,1379	−0,09	0,4641	0,09	0,5359	1,09	0,8621	2,09	0,9817
−2,10	0,0179	−1,10	0,1357	−0,10	0,4602	0,10	0,5398	1,10	0,8643	2,10	0,9821
−2,11	0,0174	−1,11	0,1335	−0,11	0,4562	0,11	0,5438	1,11	0,8665	2,11	0,9826
−2,12	0,0170	−1,12	0,1314	−0,12	0,4522	0,12	0,5478	1,12	0,8686	2,12	0,9830
−2,13	0,0166	−1,13	0,1292	−0,13	0,4483	0,13	0,5517	1,13	0,8708	2,13	0,9834
−2,14	0,0162	−1,14	0,1271	−0,14	0,4443	0,14	0,5557	1,14	0,8729	2,14	0,9838
−2,15	0,0158	−1,15	0,1251	−0,15	0,4404	0,15	0,5596	1,15	0,8749	2,15	0,9842

Продолжение табл. 7.3.1

Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность
−2,16	0,0154	−1,16	0,1230	−0,16	0,4364	0,16	0,5636	1,16	0,8770	2,16	0,9846
−2,17	0,0150	−1,17	0,1210	−0,17	0,4325	0,17	0,5675	1,17	0,8790	2,17	0,9850
−2,18	0,0146	−1,18	0,1190	−0,18	0,4286	0,18	0,5714	1,18	0,8810	2,18	0,9854
−2,19	0,0143	−1,19	0,1170	−0,19	0,4247	0,19	0,5753	1,19	0,8830	2,19	0,9857
−2,20	0,0139	−1,20	0,1151	−0,20	0,4207	0,20	0,5793	1,20	0,8849	2,20	0,9861
−2,21	0,0136	−1,21	0,1131	−0,21	0,4168	0,21	0,5832	1,21	0,8869	2,21	0,9864
−2,22	0,0132	−1,22	0,1112	−0,22	0,4129	0,22	0,5871	1,22	0,8888	2,22	0,9868
−2,23	0,0129	−1,23	0,1093	−0,23	0,4090	0,23	0,5910	1,23	0,8907	2,23	0,9871
−2,24	0,0125	−1,24	0,1075	−0,24	0,4052	0,24	0,5948	1,24	0,8925	2,24	0,9875
−2,25	0,0122	−1,25	0,1056	−0,25	0,4013	0,25	0,5987	1,25	0,8944	2,25	0,9878
−2,26	0,0119	−1,26	0,1038	−0,26	0,3974	0,26	0,6026	1,26	0,8962	2,26	0,9881
−2,27	0,0116	−1,27	0,1020	−0,27	0,3936	0,27	0,6064	1,27	0,8980	2,27	0,9884
−2,28	0,0113	−1,28	0,1003	−0,28	0,3897	0,28	0,6103	1,28	0,8997	2,28	0,9887
−2,29	0,0110	−1,29	0,0985	−0,29	0,3859	0,29	0,6141	1,29	0,9015	2,29	0,9890
−2,30	0,0107	−1,30	0,0968	−0,30	0,3821	0,30	0,6179	1,30	0,9032	2,30	0,9893
−2,31	0,0104	−1,31	0,0951	−0,31	0,3783	0,31	0,6217	1,31	0,9049	2,31	0,9896
−2,32	0,0102	−1,32	0,0934	−0,32	0,3745	0,32	0,6255	1,32	0,9066	2,32	0,9898
−2,33	0,0099	−1,33	0,0918	−0,33	0,3707	0,33	0,6293	1,33	0,9082	2,33	0,9901
−2,34	0,0096	−1,34	0,0901	−0,34	0,3669	0,34	0,6331	1,34	0,9099	2,34	0,9904
−2,35	0,0094	−1,35	0,0885	−0,35	0,3632	0,35	0,6368	1,35	0,9115	2,35	0,9906
−2,36	0,0091	−1,36	0,0869	−0,36	0,3594	0,36	0,6406	1,36	0,9131	2,36	0,9909
−2,37	0,0089	−1,37	0,0853	−0,37	0,3557	0,37	0,6443	1,37	0,9147	2,37	0,9911
−2,38	0,0087	−1,38	0,0838	−0,38	0,3520	0,38	0,6480	1,38	0,9162	2,38	0,9913
−2,39	0,0084	−1,39	0,0823	−0,39	0,3483	0,39	0,6517	1,39	0,9177	2,39	0,9916
−2,40	0,0082	−1,40	0,0808	−0,40	0,3446	0,40	0,6554	1,40	0,9192	2,40	0,9918
−2,41	0,0080	−1,41	0,0793	−0,41	0,3409	0,41	0,6591	1,41	0,9207	2,41	0,9920
−2,42	0,0078	−1,42	0,0778	−0,42	0,3372	0,42	0,6628	1,42	0,9222	2,42	0,9922
−2,43	0,0075	−1,43	0,0764	−0,43	0,3336	0,43	0,6664	1,43	0,9236	2,43	0,9925
−2,44	0,0073	−1,44	0,0749	−0,44	0,3300	0,44	0,6700	1,44	0,9251	2,44	0,9927
−2,45	0,0071	−1,45	0,0735	−0,45	0,3264	0,45	0,6736	1,45	0,9265	2,45	0,9929
−2,46	0,0069	−1,46	0,0721	−0,46	0,3228	0,46	0,6772	1,46	0,9279	2,46	0,9931
−2,47	0,0068	−1,47	0,0708	−0,47	0,3192	0,47	0,6808	1,47	0,9292	2,47	0,9932
−2,48	0,0066	−1,48	0,0694	−0,48	0,3156	0,48	0,6844	1,48	0,9306	2,48	0,9934

Продолжение табл. 7.3.1

Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность
−2,49	0,0064	−1,49	0,0681	−0,49	0,3121	0,49	0,6879	1,49	0,9319	2,49	0,9936
−2,50	0,0062	−1,50	0,0668	−0,50	0,3085	0,50	0,6915	1,50	0,9332	2,50	0,9938
−2,51	0,0060	−1,51	0,0655	−0,51	0,3050	0,51	0,6950	1,51	0,9345	2,51	0,9940
−2,52	0,0059	−1,52	0,0643	−0,52	0,3015	0,52	0,6985	1,52	0,9357	2,52	0,9941
−2,53	0,0057	−1,53	0,0630	−0,53	0,2981	0,53	0,7019	1,53	0,9370	2,53	0,9943
−2,54	0,0055	−1,54	0,0618	−0,54	0,2946	0,54	0,7054	1,54	0,9382	2,54	0,9945
−2,55	0,0054	−1,55	0,0606	−0,55	0,2912	0,55	0,7088	1,55	0,9394	2,55	0,9946
−2,56	0,0052	−1,56	0,0594	−0,56	0,2877	0,56	0,7123	1,56	0,9406	2,56	0,9948
−2,57	0,0051	−1,57	0,0582	−0,57	0,2843	0,57	0,7157	1,57	0,9418	2,57	0,9949
−2,58	0,0049	−1,58	0,0571	−0,58	0,2810	0,58	0,7190	1,58	0,9429	2,58	0,9951
−2,59	0,0048	−1,59	0,0559	−0,59	0,2776	0,59	0,7224	1,59	0,9441	2,59	0,9952
−2,60	0,0047	−1,60	0,0548	−0,60	0,2743	0,60	0,7257	1,60	0,9452	2,60	0,9953
−2,61	0,0045	−1,61	0,0537	−0,61	0,2709	0,61	0,7291	1,61	0,9463	2,61	0,9955
−2,62	0,0044	−1,62	0,0526	−0,62	0,2676	0,62	0,7324	1,62	0,9474	2,62	0,9956
−2,63	0,0043	−1,63	0,0516	−0,63	0,2643	0,63	0,7357	1,63	0,9484	2,63	0,9957
−2,64	0,0041	−1,64	0,0505	−0,64	0,2611	0,64	0,7389	1,64	0,9495	2,64	0,9959
−2,65	0,0040	−1,65	0,0495	−0,65	0,2578	0,65	0,7422	1,65	0,9505	2,65	0,9960
−2,66	0,0039	−1,66	0,0485	−0,66	0,2546	0,66	0,7454	1,66	0,9515	2,66	0,9961
−2,67	0,0038	−1,67	0,0475	−0,67	0,2514	0,67	0,7486	1,67	0,9525	2,67	0,9962
−2,68	0,0037	−1,68	0,0465	−0,68	0,2483	0,68	0,7517	1,68	0,9535	2,68	0,9963
−2,69	0,0036	−1,69	0,0455	−0,69	0,2451	0,69	0,7549	1,69	0,9545	2,69	0,9964
−2,70	0,0035	−1,70	0,0446	−0,70	0,2420	0,70	0,7580	1,70	0,9554	2,70	0,9965
−2,71	0,0034	−1,71	0,0436	−0,71	0,2389	0,71	0,7611	1,71	0,9564	2,71	0,9966
−2,72	0,0033	−1,72	0,0427	−0,72	0,2358	0,72	0,7642	1,72	0,9573	2,72	0,9967
−2,73	0,0032	−1,73	0,0418	−0,73	0,2327	0,73	0,7673	1,73	0,9582	2,73	0,9968
−2,74	0,0031	−1,74	0,0409	−0,74	0,2296	0,74	0,7704	1,74	0,9591	2,74	0,9969
−2,75	0,0030	−1,75	0,0401	−0,75	0,2266	0,75	0,7734	1,75	0,9599	2,75	0,9970
−2,76	0,0029	−1,76	0,0392	−0,76	0,2236	0,76	0,7764	1,76	0,9608	2,76	0,9971
−2,77	0,0028	−1,77	0,0384	−0,77	0,2206	0,77	0,7794	1,77	0,9616	2,77	0,9972
−2,78	0,0027	−1,78	0,0375	−0,78	0,2177	0,78	0,7823	1,78	0,9625	2,78	0,9973
−2,79	0,0026	−1,79	0,0367	−0,79	0,2148	0,79	0,7852	1,79	0,9633	2,79	0,9974
−2,80	0,0026	−1,80	0,0359	−0,80	0,2119	0,80	0,7881	1,80	0,9641	2,80	0,9974
−2,81	0,0025	−1,81	0,0351	−0,81	0,2090	0,81	0,7910	1,81	0,9649	2,81	0,9975

Окончание табл. 7.3.1

Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность	Значение z	Вероятность
–2,82	0,0024	–1,82	0,0344	–0,82	0,2061	0,82	0,7939	1,82	0,9656	2,82	0,9976
–2,83	0,0023	–1,83	0,0336	–0,83	0,2033	0,83	0,7967	1,83	0,9664	2,83	0,9977
–2,84	0,0023	–1,84	0,0329	–0,84	0,2005	0,84	0,7995	1,84	0,9671	2,84	0,9977
–2,85	0,0022	–1,85	0,0322	–0,85	0,1977	0,85	0,8023	1,85	0,9678	2,85	0,9978
–2,86	0,0021	–1,86	0,0314	–0,86	0,1949	0,86	0,8051	1,86	0,9686	2,86	0,9979
–2,87	0,0021	–1,87	0,0307	–0,87	0,1922	0,87	0,8078	1,87	0,9693	2,87	0,9979
–2,88	0,0020	–1,88	0,0301	–0,88	0,1894	0,88	0,8106	1,88	0,9699	2,88	0,9980
–2,89	0,0019	–1,89	0,0294	–0,89	0,1867	0,89	0,8133	1,89	0,9706	2,89	0,9981
–2,90	0,0019	–1,90	0,0287	–0,90	0,1841	0,90	0,8159	1,90	0,9713	2,90	0,9981
–2,91	0,0018	–1,91	0,0281	–0,91	0,1814	0,91	0,8186	1,91	0,9719	2,91	0,9982
–2,92	0,0018	–1,92	0,0274	–0,92	0,1788	0,92	0,8212	1,92	0,9726	2,92	0,9982
–2,93	0,0017	–1,93	0,0268	–0,93	0,1762	0,93	0,8238	1,93	0,9732	2,93	0,9983
–2,94	0,0016	–1,94	0,0262	–0,94	0,1736	0,94	0,8264	1,94	0,9738	2,94	0,9984
–2,95	0,0016	–1,95	0,0256	–0,95	0,1711	0,95	0,8289	1,95	0,9744	2,95	0,9984
–2,96	0,0015	–1,96	0,0250	–0,96	0,1685	0,96	0,8315	1,96	0,9750	2,96	0,9985
–2,97	0,0015	–1,97	0,0244	–0,97	0,1660	0,97	0,8340	1,97	0,9756	2,97	0,9985
–2,98	0,0014	–1,98	0,0239	–0,98	0,1635	0,98	0,8365	1,98	0,9761	2,98	0,9986
–2,99	0,0014	–1,99	0,0233	–0,99	0,1611	0,99	0,8389	1,99	0,9767	2,99	0,9986
–3,00	0,0013	–2,00	0,0228	–1,00	0,1587	1,00	0,8413	2,00	0,9772	3,00	0,9987

Решение задач на вычисление вероятности при нормальном распределении

Типичная словесно поставленная проблема, при решении которой используется нормальное распределение, представляет собой описание ситуации деловой жизни, в которой известны значения среднего и стандартного отклонения. Задача заключается в поиске одной или нескольких представляющих интерес вероятностей. Вот пример подобной словесно сформулированной задачи.

Руководство компании *Simplified Technologies, Inc.* заявило о том, что прогнозирование объемов продаж оказывается, как правило, неверным. Объем продаж за последний квартал прогнозировался на уровне $18 000 000, однако достигнутый объем составил $21 300 000. На следующий квартал прогнозируется объем продаж в $20 000 000 со стандартным отклонением (которое следует из предыдущего опыта работы) в $3 000 000. В предположении о том, что объем продаж имеет нормальное распределение с центром в прогнозируемом значении, необходимо найти вероятность того, что следующий квартал окажется "действительно неудачным", чему соответствует объем продаж, меньший $15 000 000.

Начало приведенного рассказа описывает сцену, на которой разворачивается действие. Первые приведенные числа (18 и 21,3) описывают предыдущие события и не играют роли в нашей задаче. Внимание следует обратить на следующие факты:

В рассматриваемой задаче присутствует нормальное распределение.

Его среднее значение μ = $20 000 000.

Его стандартное отклонение σ = $3 000 000.

Необходимо найти вероятность того, что объем продаж окажется ниже $15 000 000.

Следующий шаг состоит в том, чтобы пронормировать все эти значения (за исключением среднего и стандартного отклонения), что позволит использовать для поиска ответа таблицу вероятностей для стандартного распределения. **Нормированное значение** (часто его обозначают z) представляет собой число стандартных отклонений от среднего в большую (если нормированное значение положительное) или меньшую (если нормированное значение отрицательное) сторону. Это преобразование выполняется следующим образом:

$$\text{z = Нормированное значение} = \frac{\text{Значение – Среднее}}{\text{Стандартное отклонение}} = \frac{\text{Значение} - \mu}{\sigma}.$$

В приведенном примере значение $15 000 000 нормируется следующим образом:

$$z = \frac{15 - \mu}{\sigma} = \frac{15 - 20}{3} = -1{,}67.$$

Это означает, что в данном случае величина $15 000 000 соответствует z = –1,67 стандартным отклонениям ниже среднего (прогнозируемого) значения.[9] Таким образом, мы преобразовали исходную задачу в следующую задачу вычисления вероятности для стандартного нормального распределения.

Найти вероятность того, что имеющая стандартное нормальное распределение величина окажется меньше z = –1,67.

По таблице находим ответ на поставленный вопрос.

Вероятность того, что квартал окажется действительно неудачным, составляет 0,0475, или приблизительно 5%.

Вот и хорошо. Похоже, что вероятность действительно неудачного квартала не очень велика. Однако вероятность 5% тоже не следует недооценивать.

Рис. 7.3.6 и 7.3.7 иллюстрируют вычисление вероятности в терминах объемов продаж в долларах и в терминах нормированных значений (количествах стандартных отклонений от среднего в сторону больших или меньших значений).

Эта задача оказывается достаточно простой, поскольку ответ можно найти непосредственно из таблицы вероятностей для стандартного нормального распределения. Ответ на вопрос, приведенный ниже, требует больших усилий.

[9] Речь идет об отклонении ниже среднего, поскольку нормированное значение z = –1,67 — отрицательное. При любом превышении среднего нормированное значение z будет положительным. Для среднего нормированное значение z равно 0.

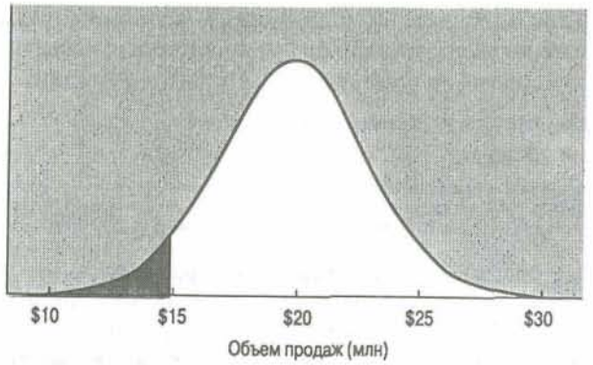


Рис. 7.3.6. Вероятность того, что квартал действительно окажется неудачным (объем продаж меньше $15 000 000), представлена заштрихованной областью под кривой. Результат основан на прогнозируемом объеме продаж $20 000 000 и стандартном отклонении $3 000 000. Задача решена нормированием и последующим применением таблицы вероятностей для стандартного нормального распределения

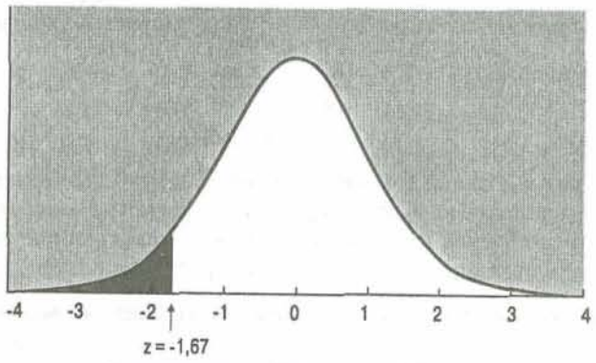


Рис. 7.3.7. Вероятность того, что квартал окажется действительно неудачным, представленная с использованием нормированного значения объема продаж. Это вероятность отклонения объема продаж от среднего более чем на z = −1,67 величин стандартного отклонения в сторону уменьшения. Результат равен 0,0475

Для описанной выше задачи о прогнозировании объема продаж найти вероятность "действительно удачного квартала", который определяется как такой, в котором объем продаж превышает $24 000 000.

Первый шаг на пути к ответу состоит в нормировании объема продаж: значению $24 000 000 соответствует z = (24 − 20)/3 = 1,33 величины стандартного отклонения от среднего в сторону превышения. Таким образом, необходимо дать ответ на следующий вопрос.

Найти вероятность того, что имеющая стандартное нормальное распределение случайная величина окажется *больше* z = 1,33.

Из правила дополнительности известно, что необходимая нам вероятность равна единице минус вероятность того, что случайная величина окажется меньше z=1,33. Находим в таблице вероятность, соответствующую 1,33, и вычисляем результат.

Вероятность действительно удачного квартала = 1 − 0,9082 =
= 0,0918, или примерно 9%.

Эта вероятность с использованием нормированных значений показана на рис. 7.3.8.

Рассмотрим еще одну задачу.

Для описанной выше задачи на прогнозирование объема продаж найти вероятность "типичного квартала", который определяется как соответствующий объему продаж от $16 000 000 до $23 000 000.

Прежде всего пронормируем оба приведенных в условии значения. После этого задача примет следующий вид.

Найти вероятность того, что имеющая стандартное нормальное распределение величина принимает значения *между* $z_1 = -1,33$ и $z_2 = 1,00$.

Для того чтобы найти ответ на поставленный в условии этой задачи вопрос, необходимо найти в таблице вероятность, соответствующую каждому из этих нормированных значений, а затем найти разность этих значений. Необходимо следить за тем, чтобы вычиталось меньшее значение из большего, а не наоборот — тогда результат будет положительным и его действительно можно будет интерпретировать как вероятность.

Вероятность типичного квартала = 0,8413 − 0,0918 =
= 0,7495, или примерно 75%.

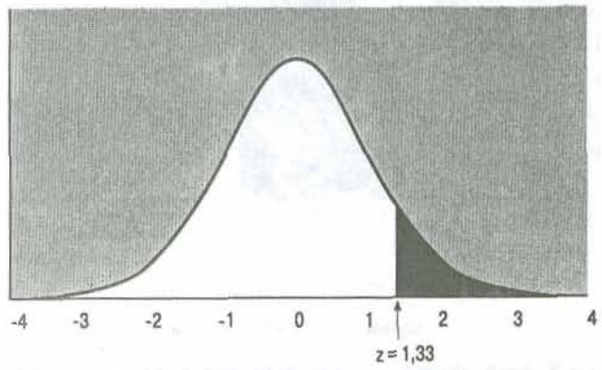


Рис. 7.3.8. Вероятность того, что квартал окажется действительно удачным, представленная в терминах нормированных значений объема продаж. Заштрихованная область равна 1 минус незаштрихованная область под кривой. Число, соответствующее незаштрихованной области, можно найти в таблице. Результат равен 0,0918

Эта вероятность с использованием нормированных значений показана на рис. 7.3.9.

И, наконец, еще одна, несколько отличающаяся от предыдущих, задача.

Для описанной выше задачи о прогнозировании объема продаж найти вероятность "необычного квартала", который определяется как соответствующий объему продаж, либо меньших $16 000 000, либо больших $23 000 000.

В этом случае речь идет о вероятности *не* попасть между двумя указанными значениями. Если воспользоваться правилом дополнительности, чтобы найти ответ, можно просто вычесть из единицы вероятность, вычисленную в предыдущем случае, — вероятность того, что объем продаж попадет в область между этими же двумя числами. При этом получаем следующий результат:

$$\text{Вероятность необычного квартала} = 1 - 0{,}7495 = 0{,}2505, \text{ или примерно } 25\%.$$

Эта вероятность с использованием нормированных значений показана на рис. 7.3.10.

Для вычисления значений вероятности из наших первых трех примеров в Excel® используется функция `=NORMDIST(value, mean, standardDeviation, TRUE)` (`=НОРМРАСП(Значение; Среднее; СтандОткл; Истина)`). Эта функция дает возможность найти вероятность того, что величина, имеющая нормальное распределение с некоторым средним значением (`mean`) и стандартным отклонением (`standardDeviation`), окажется меньше некоторого значения (`value`). Нормировать значение (`value`) нет необходимости, поскольку Excel делает это самостоятельно. Вычисление вероятности из первого примера производится не-

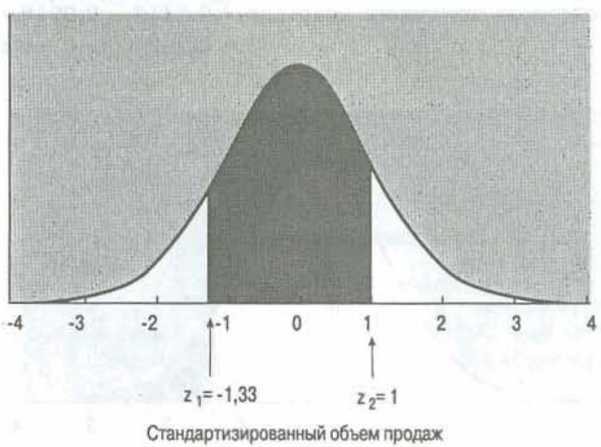


Рис. 7.3.9. Вероятность того, что квартал окажется типичным, представленная в терминах нормированных значений объема продаж. Число, соответствующее заштрихованной области, можно найти как разность чисел, определяемых по таблице для каждого из нормированных значений. При вычитании из результата удаляется левый незаштрихованный *участок. Результат равен 0,7495*

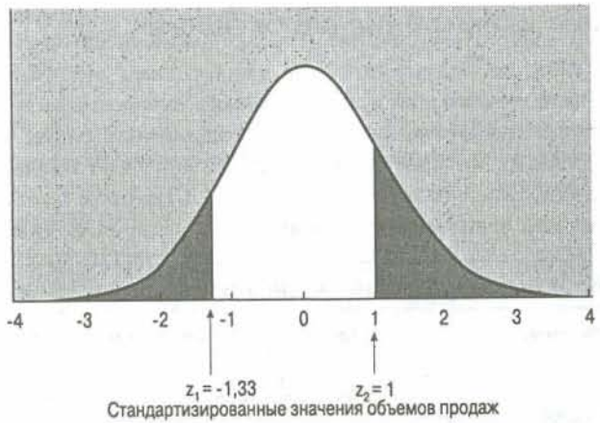
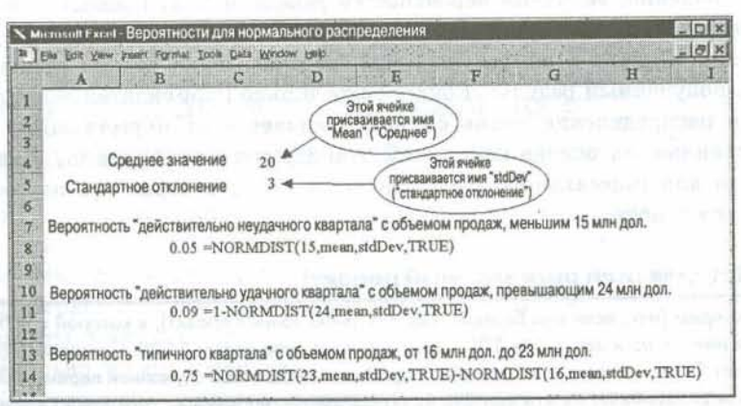


Рис. 7.3.10. Вероятность того, что квартал окажется необычным, представленная в терминах нормированных значений объема продаж. Число, соответствующее заштрихованной области, можно найти в результате вычитания из единицы разности величин вероятностей, определяемых по таблице для каждого из нормированных значений. Результат равен 0,2505

посредственно, поскольку нас интересует как раз вероятность того, что случайная величина меньше некоторого значения. Во втором примере вероятность вычисляется как единица минус результат функции `NORMDIST` (`НОРМРАСП`) — при этом получается вероятность того, что случайная величина превышает некоторое значение. В третьем примере вычисляется разность двух значений функции `NORMDIST` (`НОРМРАСП`), поскольку необходимо определить вероятность того, что объем продаж попадет в интервал между двумя определенными значениями. На приведенном ниже рисунке показано применение Excel для решения представленных выше задач.



Четыре способа вычисления вероятности

В следующей таблице кратко описаны четыре типа задач на вычисление вероятности и соответствующие способы решения. Значения z, z_1 и z_2 — это нормированные значения из постановок задач, полученные путем вычитания среднего из представляющих интерес значений и последующего деления полученного результата на стандартное отклонение. Таблица, о которой идет речь, — это таблица вероятностей для стандартного нормального распределения.

Вычисление вероятностей в случае нормального распределения

Найти вероятность того, что величина Z...	Необходимые действия
Меньше z	Найти в таблице значение вероятности для z
Больше z	Вычесть из 1 предыдущий результат
Между z_1 и z_2	Найти в таблице вероятности, соответствующие z_1 и z_2, и вычесть из большего значения меньшее
За пределами интервала между z_1 и z_2	Вычесть из 1 предыдущий (для вопроса "Между z_1 и z_2") результат

Могут возникнуть некоторые сомнения относительно того, существует ли различие между событиями "объем продаж превышает $22 000 000" и "объем продаж на уровне не менее $22 000 000". Условие *превышает* означает *более чем*, в то время как *не менее* означает *больше или равно*. В случае нормального распределения вероятности таких двух событий практически не отличаются. Это связано с тем, что различие между двумя такими вероятностями представляется только вертикальным отрезком линии, который не ограничивает никакой площади под кривой нормального распределения.

Внимание! Не все распределено нормально!

В случае нормального распределения, если известны среднее значение и нормальное отклонение, значения вероятностей можно найти, проведя нормировку и воспользовавшись затем таблицей вероятностей для стандартного нормального распределения. Однако в том случае, если распределение только приблизительно нормально, полученный результат будет тоже только приблизительно правильным.

Если же распределение очень сильно отличается от нормального, вероятности, вычисленные на основе среднего и стандартного отклонения, с использованием таблиц для нормального распределения могут совершенно не соответствовать действительности.

Пример. *Лотерея (или рискованный проект)*

Рассмотрим лотерею (или, если вам больше нравится, рискованный проект), в которой в 90% случаев не выплачивается ничего, но в остальных 10% случаев выплата составляет $500. Ожидаемая (средняя) выплата составляет $50, а стандартное отклонение для такой дискретной случайной переменной составляет $150. Обратите внимание на то, что выплаты не распределены нормально; отсутствует даже отдаленное сходство с нормальным распределением, поскольку возможные значения выплаты дискретны и могут принимать только одно из двух значений.

Чему равна вероятность выигрыша по меньшей мере $50? Правильный ответ — 10%, поскольку единственный способ что-либо выиграть состоит в том, чтобы выиграть полную сумму, $500, которая и содержит "по меньшей мере $50".

Попробуем рассмотреть нормальное распределение с такими же значениями среднего ($50) и стандартного отклонения ($150). Насколько далеким от правильного ответа (10%) окажется оценка вероятности выигрыша, если исходить из свойств нормального распределения? Отличие будет очень большим, поскольку вероятность того, что следующая нормальному распределению случайная величина примет значение, превышающее среднее, равна 0,5, или 50%.

Это действительно существенное различие: 10% (правильный ответ) и 50% (результат вычислений при неверном предположении, что случайная величина имеет нормальное распределение). На рис. 7.3.11 показано различие между реальным дискретным распределением и нормальным распределением с таким же средним значением и стандартным отклонением. Выдвигая предположение о том, что величина имеет нормальное распределение, нужно всегда быть очень осторожным.

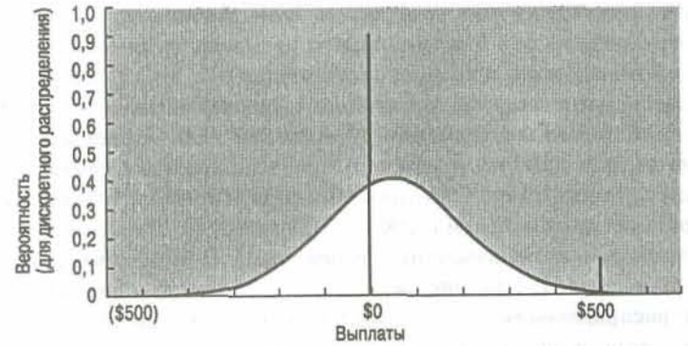


Рис. 7.3.11. Дискретное распределение размера выплаты и нормальное распределение с таким же средним значением ($50) и стандартным отклонением ($150). Эти распределения и соответствующие вероятности сильно различаются между собой. Дискретное распределение дает правильный ответ; предположение о том, что размеры выплат распределены нормально, в данном случае неверно

7.4. Аппроксимация биномиального распределения нормальным

Давайте вспомним: биномиальное распределение описывает количество наступлений некоторого события в n независимых попытках. Биномиальное распределение никогда не может в точности совпадать с нормальным в силу двух причин. Во-первых, любое нормальное распределение может давать наблюдаемые результаты в виде чисел с дробной частью (например, 7,11327), в то время как биномиально распределенная величина X может принимать только целые значения (допустимым является, например, число 7). Кроме того, биномиальное распределение при π, отличном от 0,5, всегда асимметрично, в то время как нормальное распределение во всех случаях сохраняет идеальную симметрию.

Однако биномиальное распределение можно хорошо аппроксимировать с помощью нормального распределения, если n достаточно велико, а вероятность π не слишком близка к 0 или 1.[10, 11] Это помогает вычислять вероятности (того, что некоторая величина меньше определенного значения, превышает его, находится между двумя значениями или вне интервала между двумя значениями) для биномиального распределения путем замены многих сложных и трудоемких вычислений (по рассмотренной ранее формуле для вычисления вероятностей, имеющих биномиальное распределение величин) на более простые вычисления (с использованием формул для нормального распределения).

Как, однако, выбрать такое нормальное распределение, которое достаточно близко к данному биномиальному распределению? Хорошим выбором будет использование нормального распределения с такими же значениями среднего и стандартного отклонения, как и у подлежащего аппроксимации биномиального распределения. Поскольку мы уже знаем, как вычислять среднее значение и стандартное отклонение для биномиального распределения (этот вопрос рассмотрен в разделе 7.2), а также как вычислять вероятности для нормального распределения с известными средним и стандартным отклонениями (см. раздел 7.3), вычисление приближенных значений вероятности для имеющих биномиальное распределение величин уже не должно представлять особых сложностей.

Рассмотрим пример аппроксимации биномиального распределения нормальным. Предположим, что n равно 100 и π равно 0,10. Распределение вероятностей, вычисленных с использованием формулы для биномиального распределения, показано на рис. 7.4.1. Распределение достаточно явно имеет присущую нормальному распределению колоколообразную форму. Несмотря на то что распределение все еще остается дискретным, достаточно очевидно, что эта дискретность не является его главным свойством.

Для того чтобы аппроксимировать биномиальное распределение (с дискретными целочисленными значениями) с помощью нормально распределенной случайной величины (непрерывной), отложим от каждого значения вправо и влево 1/2, чтобы включить в рассмотрение все числа, расположенные вокруг целых чисел.[12] Например, для аппроксимации вероятности того, что некоторая биномиально распределенная величина X равна 3, необходимо найти вероятность того, что нормально распределенная (с теми же значениями среднего и стандартного отклонения) величина попадает в интервал от 2,5 до 3,5. Такое расширение

[10] Если π близко к 0 или 1, приближение к нормальному распределению с ростом n оказывается более медленным, что обусловлено асимметрией биномиального распределения с редкими или почти определенными событиями. Хорошим приближением для биномиального распределения при больших n и близких к 0 значениях π оказывается распределение Пуассона, которое будет рассмотрено в следующем разделе.

[11] Центральная предельная теорема, которую мы рассмотрим в главе 8, поясняет возникновение нормального распределения при объединении результатов большого числа независимых случайных попыток.

[12] Здесь предполагается поиск вероятностей для имеющего биномиальное распределение количества X наступлений события. Если необходимо вычислить вероятности биномиальной доли, или процента, p, следует сначала перейти к количеству X. Например, вероятность наблюдения "по меньшей мере 20% из 261" — это то же, что и наблюдение "по меньшей мере 53 из 261", поскольку в этом случае требуется по меньшей мере $0{,}20 \times 261 = 52{,}2$ наблюдений, причем возможны только целые значения.

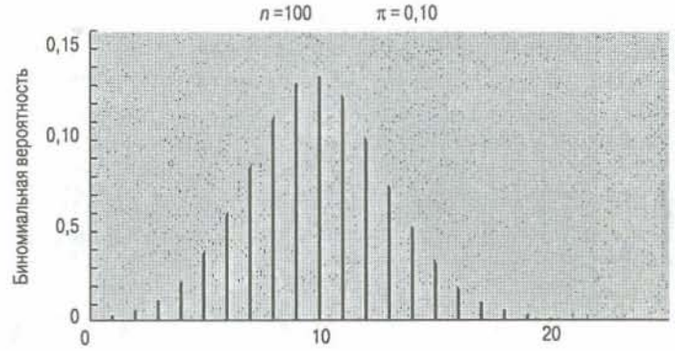


Рис. 7.4.1. Биномиальное распределение при n =100 и π = 0,10 достаточно близко к нормальному

необходимо в связи с тем, что для любой нормально распределенной случайной величины вероятность ее точного равенства числу 3 равна нулю, и в то же время все значения нормально распределенной случайной величины в интервале от 2,5 до 3,5 округляются до целого числа 3. Аналогичным образом вероятность того, что биномиально распределенная случайная величина примет значение в интервале от 6 до 9, соответствует вероятности того, что нормально распределенная (с тем же значениями среднего и стандартного отклонения) величина попадает в промежуток от 5,5 до 9,5. Вероятность того, что значение окажется *вне* ограниченного двумя числами интервала, равна, как обычно, единице минус вероятность попадания в этот интервал.

Аппроксимация биномиального распределения нормальным (a и b — целые)

Вероятность того, что биномиально распределенная величина...	Аппроксимируется вероятностью того, что соответствующая нормально распределенная величина принимает значения ...
равна 8	от 7,5 до 8,5
равна a	от a–0,5 до a+0,5
лежит между 15 и 23	от 14,5 до 23,5
лежит между a и b	от a–0,5 до b+0,5

Сравните графики, показанные на рис. 7.4.1 (n = 100) и рис. 7.4.2 (n = 10). Из них видно, что при меньших значениях n распределение менее похоже на нормальное. Кроме того, при меньших n больше сказывается дискретность распределения.

Пример. *Быстрые и медленные микропроцессоры*

Производственный процесс часто контролируется не так хорошо, как хотелось бы. Это утверждение относится и к производству используемых в микрокомпьютерах сложных микропроцессорных интегральных микросхем, в которых более 1 000 000 транзисторов размещаются на кремниевой подложке площадью в 1/4 квадратного дюйма. Несмотря на тщательный контроль, получаемые микросхемы отличаются друг от друга: одни оказываются более быстродействующими, чем другие.

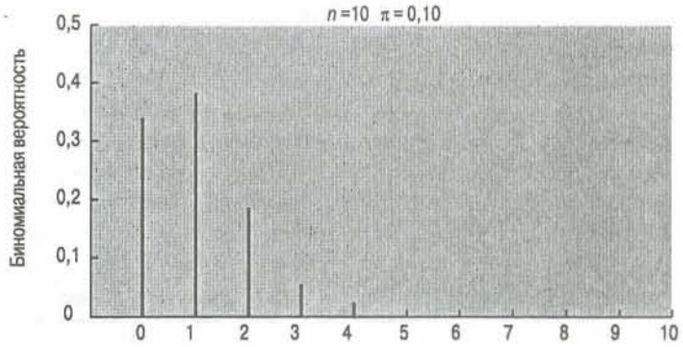


Рис. 7.4.2. Биномиальное распределение при n = 10 и π = 0,10 не очень близко к нормальному, поскольку n недостаточно велико

В традициях программного обеспечения предыдущих поколений, когда утверждалось: "Это не ошибка, это такая особенность!", произведенные микросхемы сортируются по быстродействию и цены на них устанавливаются соответствующим образом (более быстрые схемы дороже и стоят). В каталоге указывают два типа изделий: со скоростью 300 мегагерц (более медленные) и 500 мегагерц (более быстрые).

Используемое некоторой компанией оборудование в 80% случаев производит медленные микросхемы, а быстрые микросхемы составляют остальные 20% объема выпускаемой продукции. Каждая схема оказывается быстрой или медленной независимо от свойств схем, произведенных до и после нее. Представим себе, что сегодня нужно отгрузить 1000 медленных микросхем и 300 быстрых микросхем; возможно, при этом какие-то микросхемы останутся в избытке. Выпуск какого количества микросхем необходимо запланировать?

Если в производственный план внести выпуск 1300 микросхем, следует ожидать, что 80% (1040 микросхем) окажутся медленными, а 20% (260 микросхем) будут быстродействующими. Медленных схем окажется достаточно, однако, в среднем быстродействующих будет мало.

Поскольку очевидно, что минимально необходимый объем плана выпуска определяется быстродействующими микросхемами, на их количестве и следует основывать расчеты. Сначала находим 300/0,20 = 1500. Отсюда следует, что если в план внести выпуск 1500 микросхем, можно ожидать, что 20% из них (300 штук) окажутся быстродействующими. Это позволит в среднем достичь соответствия поставленной цели. Однако, к сожалению, при этом вероятность выполнения поставленной задачи по выпуску быстродействующих микросхем составит только около 50%!

Предположим теперь, что запланирован выпуск 1650 микросхем. Чему равна вероятность того, что цель будет достигнута? Для получения ответа на этот вопрос прежде всего сформулируем его как задачу на вычисление вероятности.

> Дана биномиально распределенная случайная величина (количество выпущенных быстродействующих микросхем) с общим количеством микросхем n =1650 и вероятностью того, что микросхема окажется быстродействующей, π = 0,20. Необходимо найти вероятность того, что эта случайная величина примет значение, равное по меньшей мере 300, но не превышающее 650.[13]

Если решать эту задачу, непосредственно вычисляя вероятность для биномиального распределения, придется рассчитывать вероятности для 300, 301, 302 и т.д. схем. Аппроксимирование биномиального распределения нормальным позволяет получить ответ значительно быстрее с помощью таблицы стандартного нормального распределения. При этом необходимо знать среднее значение и стандартное отклонение для количества произведенных быстрых микросхем.

$$\mu_{\text{(количество быстрых микросхем)}} = n\pi = 1650 \times 0{,}20 = 330$$

$$\sigma_{\text{(количество быстрых микросхем)}} = \sqrt{n\pi(1-\pi)} = \sqrt{1650 \times 0{,}20 \times 0{,}80} = 16{,}24807.$$

Необходимо также нормировать предельные значения для количества требуемых быстрых микросхем, 300 и 650 (после расширения интервала на S получаем 299,5 и 650,5). Нормирование проводится с использованием уже найденных среднего значения и стандартного отклонения:

$$z_1 = \text{Нормированный нижний предел количества быстрых схем} = \frac{299{,}5-330}{16{,}24807} = -1{,}88$$

$$z_2 = \text{Нормированный верхний предел количества быстрых схем} = \frac{650{,}5-330}{16{,}24807} = 19{,}73.$$

Соответствующие этим нормированным величинам значения вероятности находим в таблице стандартного нормального распределения. Для $z_1 = -1{,}88$ это 0,030. Поскольку число $z_2 = 19{,}73$ лежит за пределами таблицы, соответствующую ему вероятность принимаем равной 1.[14] Вычитая меньшую вероятность из большей, находим вероятность того, что случайная величина будет лежать в указанных пределах, и таким образом получаем необходимый ответ: 1 – 0,030 = 0,970. Отсюда можно сделать вывод о том, что если в производственный план включить выпуск 1650 микросхем, то с вероятностью 97% цель отгрузить 300 быстродействующих микросхем и 1000 медленных микросхем будет достигнута.

Вероятности помогают также понять, что происходит "за кулисами" действия, разворачивающегося в реальной жизни. Попробуем разобраться, что может происходить при проведении социологического исследования. Воспользуемся для этого анализом сценариев вида *что если...* .

Пример. *Социологический опрос избирателей*

Фирма, специализирующаяся на социологических исследованиях и проведении опросов по телефону, получила заказ на опрос общественного мнения для выяснения того, будет ли новая инициатива местных властей поддержана при голосовании во время следующих выборов. Фирма принимает решение опросить 800 выбранных случайным образом человек, которые, видимо, примут участие в голосовании. В результате опроса установлено, что 437 человек собираются голосовать "за". Вот теперь и возникает вопрос "Что если?", который в данном случае формулируется так: если бы мнения всех избирателей разделились поровну между "за" и "против", с какой вероятностью можно было бы ожидать, что именно

[13] Это ограничение обусловлено тем, что выпуск более чем 1650 – 1000 = 650 быстрых микросхем означает, что выпущено менее чем 1000 медленных микросхем; в этом случае не удастся достичь поставленной цели с точки зрения количества медленных микросхем.

[14] И это действительно так: вероятность того, что имеющая стандартное нормальное распределение случайная величина окажется меньше, чем значение, превышающее среднее на $z_2 = 19{,}73$ стандартных отклонений, действительно, в сущности, равна 1, поскольку такое происходит практически всегда.

столько или более людей, попавших в выборку для опроса, ответили бы, что они собираются голосовать "за"? Вы ищете ответ на этот вопрос вместе со своим сотрудником.

- Ваш сотрудник: "Эти доли, похоже, достаточно близки: 437 из 800 очень близко к распределению голосов 50 на 50, что соответствовало бы 400 из 800".
- Вы: "А мне кажется, что 437 гораздо больше, чем 400. Нужно попробовать выяснить, можно ли дополнительные 37 голосов "за" объяснить только случайностью".
- Ваш сотрудник: "Хорошо. Можно предположить, что каждый из опрошенных с одинаковой вероятностью может быть "за" или "против". Тогда можно рассчитать вероятность того, что результат "за" составит 437 или более".
- Вы: "Это можно. Если вероятность окажется больше 5 или 10%, дополнительные 37 ответов "за" можно будет считать случайными. Но если вероятность будет мала, например меньше 5% или даже меньше 1%, то, видимо, здесь присутствует нечто большее, чем просто случайность".

Для того чтобы произвести соответствующие вычисления, предположим, что некоторая величина X описывает следующую биномиально распределенную случайную величину: количество людей (из 800 опрошенных), сказавших, что они собираются голосовать "за". Если предположить, что мнения по этому вопросу разделились поровну, вероятность того, что каждый из опрошенных ответит "я — за", равна $\pi = 0{,}50$. Найдем теперь среднее значение и стандартное отклонение величины X, воспользовавшись для этого соответствующими формулами для биномиального распределения:

$$\mu_X = n\pi = (800)(0{,}50) = 400$$

$$\sigma_X = \sqrt{n\pi(1-\pi)} =$$

$$= \sqrt{(800)(0{,}50)(1-0{,}50)} =$$

$$= 14{,}14214.$$

Теперь, чтобы найти вероятность того, что X принимает значение, равное по меньшей мере 437, увеличим пределы на 1/2 — при этом надо будет найти вероятность того, что X составляет по меньшей мере 436,5, и можно будет воспользоваться тем, что распределение X приблизительно нормальное. Итак, нужно найти вероятность того, что нормально распределенная случайная величина со средним значением 400 и стандартным отклонением 14,14214 превышает значение 436,5. Для этого нормируем значения:

$$z = \text{Нормированное значение} = \frac{436{,}5-\mu_x}{\sigma_x} =$$

$$= \frac{436{,}5-400}{14{,}14214} =$$

$$= 2{,}58.$$

В таблице нормального распределения находим, что в предположении равного распределения мнений среди населения вероятность того, что интересующая нас величина достигает для рассматриваемой выборки граничного значения (или превышает его), равна 1 – 0,995 = 0,005. Правдоподобие получения такого результата очень мало: вероятность составляет всего лишь половину процента, что соответствует 1 шансу из 200.

Вы задали вопрос, что будет, если мнения по интересующей вас проблеме разделились среди населения поровну, и получили на него ответ: "В таком случае получить в выборке результат 54,6% (это 437/800) или более очень маловероятно". Таким образом, использование сценария *Что если?* дало возможность опровергнуть предположение о равном распределении голосов "за" и "против" среди избирателей. Это неплохо для начала!

7.5. Распределение Пуассона и экспоненциальное распределение

Существует много других распределений вероятности, которые полезны в статистических исследованиях. В этом разделе кратко описаны два таких распределения и показано, как их можно применять в конкретных ситуациях деловой жизни.

Распределение Пуассона

Распределение Пуассона, подобно биномиальному распределению, связано с подсчетом количества наступления некоторого события. Отличие состоит в том, что в случае распределения Пуассона нет заданного числа возможных попыток n. Вот один из примеров возникновения такой случайной величины. Если некоторое событие происходит случайно и независимо в каждой из попыток и среднее число наступлений события с ростом числа попыток не изменяется, то количество наступлений события в фиксированном количестве попыток будет подчиняться **распределению Пуассона**[15]. Распределение Пуассона — это распределение дискретной величины, которое зависит только от ожидаемого среднего количества наступлений события.

Приведем примеры некоторых случайных величин, которые могут иметь распределение Пуассона.

1. Количество заказов, которые фирма получит завтра.
2. Количество людей, которые обратятся завтра в отдел кадров компании.
3. Количество дефектов в произведенной продукции.
4. Количество звонков в фирму в течение следующей недели просьбой помочь разобраться с "простой в сборке" игрушкой.
5. Биномиально распределенная величина X при больших n и малых π.

На приведенных ниже рисунках показано распределение вероятностей случайных величин, имеющих распределение Пуассона, при ожидании в среднем 0,5 наступлений соответствующего случайной величине события (рис. 7.5.1), ожидании 2 наступлений событий (рис. 7.5.2) и ожидании 20 наступлений события (рис. 7.5.3). Обратите внимание на то, что форма распределения Пуассона, показанная на рис 7.5.3, подобна колоколообразной форме нормального распределения. Это свидетельствует о том, что в случае ожидания наступления большого количества событий распределение Пуассона приближается к нормальному.

Распределение Пуассона имеет три существенные особенности, знание которых позволяет находить вероятности, если известно только среднее значение случайной величины.

Для распределения Пуассона

1. Стандартное отклонение всегда равно корню квадратному из среднего значения.
2. Вероятность того, что имеющая распределение Пуассона случайная величина X со средним значением μ равна a, выражается формулой

[15] Это распределение носит имя французского ученого Пуассона.

$$P(X=a)=e^{-\mu}\left(\frac{\mu^a}{a!}\right),$$

где e = 2,71828....

3. При больших средних значениях распределение Пуассона близко к нормальному распределению.



Рис. 7.5.1. Распределение Пуассона с ожидаемым количеством наступлений события 0,5 асимметрично. Существует большая вероятность (0,607) того, что событие вообще не наступит



Рис. 7.5.2. Вероятности для распределения Пуассона с ожидаемым количеством наступлений события, равным 2. Это распределение все еще продолжает быть несколько асимметричным

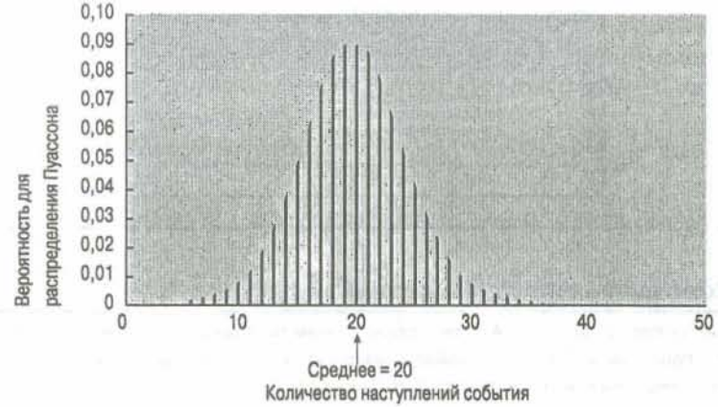


Рис. 7.5.3. Вероятности для распределения Пуассона с ожидаемым количеством наступлений события, равным 20. Это распределение продолжает оставаться дискретным, но по форме приближается к нормальному распределению

Пример. Количество возвратов товара по гарантии

Фирма работает с товарами очень высокого качества, благодаря чему каждый день ожидается возврат на гарантийный ремонт (в среднем) только 1,3 единицы товара. С какой вероятностью завтра в гарантийный ремонт не поступит ни одного изделия? Какова вероятность возврата одного изделия? Двух? Трех?

Поскольку среднее значение (1,3) очень мало, вероятности необходимо вычислять с использованием точной формулы для распределения Пуассона. Вот эти вычисления:

$$P(X=0)=e^{-1,3}\times\frac{1,3^0}{0!}=0,27253\times\frac{1}{1}=0,27253;$$

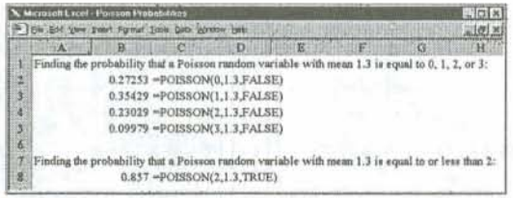
$$P(X=1)=e^{-1,3}\times\frac{1,3^1}{1!}=0,27253\times\frac{1,3}{1}=0,35429;$$

$$P(X=2)=e^{-1,3}\times\frac{1,3^2}{2!}=0,27253\times\frac{1,69}{2}=0,23029;$$

$$P(X=3)=e^{-1,3}\times\frac{1,3^3}{3!}=0,27253\times\frac{2,197}{6}=0,09979.$$

Зная эти основные вероятности, можно сложить вероятности возврата 0, 1 и 2 изделий, чтобы вычислить вероятность того, что в гарантийный ремонт поступят 2 *или менее* изделий. Вероятность такого события равна: 0,27253 + 0,35429 + 0,23029 = 0,857, или 87,5%.

Для вычисления этих вероятностей с использованием Excel применяется функция `=POISSON(value, mean, FALSE)` (`=ПУАССОН(значение; среднее; ЛОЖЬ)`), которая вычисляет вероятность того, что случайная переменная, имеющая распределение Пуассона со средним значением `mean`, принимает некоторое конкретное значение `value`, а также функция `=POISSON(value, mean, TRUE)` (`=ПУАССОН(значение; среднее; ИСТИНА)`), вычисляющая вероятность того, что значение имеющей распределение Пуассона случайной переменной будет *меньше или равно* значению `value`. Ниже приведен пример соответствующих вычислений.



	A	B	C	D	E	F	G	H
1	Finding the probability that a Poisson random variable with mean 1.3 is equal to 0, 1, 2, or 3:							
2		0.27253	=POISSON(0,1.3,FALSE)					
3		0.35429	=POISSON(1,1.3,FALSE)					
4		0.23029	=POISSON(2,1.3,FALSE)					
5		0.09979	=POISSON(3,1.3,FALSE)					
6								
7	Finding the probability that a Poisson random variable with mean 1.3 is equal to or less than 2:							
8		0.857	=POISSON(2,1.3,TRUE)					

Пример. *Количество телефонных звонков*

В среднем в фирму поступает в день 460 телефонных звонков. В предположении, что количество звонков подчиняется распределению Пуассона, найдем вероятность того, что завтрашний день окажется перегруженным, т.е. телефонных звонков окажется 500 или более.

Среднее значение дано в условии. Стандартное отклонение составляет $\sqrt{460} = 21{,}44761$. Поскольку среднее значение достаточно велико, для данного распределения можно в качестве приближения использовать нормальное распределение. Нормальное распределение — непрерывное, любое значение, превышающее 499,5, — будет округляться до числа 500 и более. Нормированное количество обращений равно:

$$z = \frac{499{,}5 - 460}{21{,}44761} = 1{,}84 .$$

Воспользовавшись таблицей стандартного нормального распределения вычисляем искомую 1 – 0,967 = 0,033. Таким образом, вероятность того, что завтрашний день окажется перегруженным, составляет всего лишь около 3% (т.е. такое событие не очень правдоподобно).

Экспоненциальное распределение

Экспоненциальное распределение — это непрерывное распределение с сильной асимметрией (рис. 7.5.4). В левой части кривая распределения при приближении к 0 уходит вертикально вверх, а в правой части постепенно понижается.

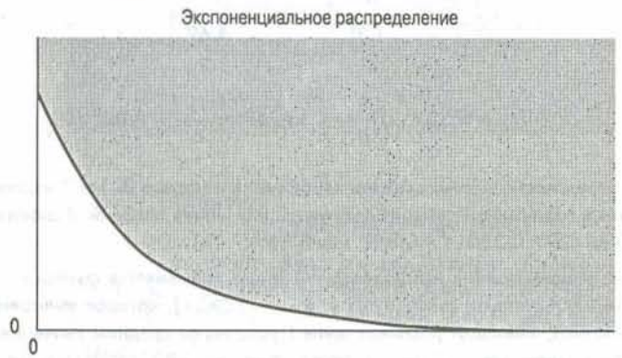


Рис. 7.5.4. Экспоненциальное распределение имеет сильную асимметрию и часто используется для представления времени ожидания между событиями

Ниже описаны случаи, к которым применимо экспоненциальное распределение. Если события происходят случайно, независимо и с постоянной частотой, время ожидания между двумя последовательно наступающими событиями имеет экспоненциальное распределение[16].

Вот несколько примеров случайных величин, которые могут иметь экспоненциальное распределение.

1. Промежутки времени между появлением посетителей в авторемонтной мастерской.
2. Периоды времени нормальной работы копировального аппарата между появлениями неисправностей, требующих вмешательства специалиста по ремонту.
3. Длительность типичного телефонного разговора.
4. Время до выхода из строя кинескопа телевизора.
5. Затраты времени на обслуживание одного покупателя.

Экспоненциальное распределение *не обладает памятью* в том несколько удивительном смысле, что если вы ожидаете события в течение некоторого промежутка времени и это событие не наступило, то в результате среднее время ожидания этого события не уменьшилось по сравнению с тем, что было в момент начала ожидания. Этот факт становится понятным, если мы вспомним, что события наступают независимо друг от друга и каждое такое событие "не знает" о том, что в последнее время предыдущее событие не происходило.

Что можно сказать, исходя из изложенного выше, о телефонных разговорах? Представьте себе, что вы руководите автоматической телефонной станцией, через которую осуществляются соединения средней длительностью пять минут. Рассмотрим все телефонные переговоры, которые начались в настоящий момент. Вы ожидаете, что эти разговоры будут продолжаться в среднем в течение пяти минут, причем длительность отдельных разговоров имеет экспоненциальное распределение. По истечении одной минуты некоторые разговоры оканчиваются. Однако для всех остальных разговоров ожидается продолжение в среднем также в течение *следующих пяти минут*. Это связано с тем, что самые короткие звонки уже удалены из рассмотрения. В такое, возможно, трудно поверить, однако это подтверждается (приблизительно) реальными данными.

Ниже описаны основные свойства экспоненциального распределения. Обратите внимание на то, что для этого распределения нет "нормальной приближения", поскольку экспоненциальное распределение *всегда* очень асимметрично.

Для экспоненциального распределения

1. Стандартное отклонение всегда равно среднему значению.
2. Вероятность того, что имеющая экспоненциальное распределение случайная величина X со средним значением μ принимает значение, меньшее a, выражается формулой

$$P(X < a) = 1 - e^{-a/\mu} .$$

[16] Обратите внимание на то, что при этом общее количество событий подчиняется распределению Пуассона.

Если события происходят независимо друг от друга и с постоянной частотой, между экспоненциальным распределением и распределением Пуассона существует определенная взаимосвязь. *Количество событий* для любого фиксированного промежутка времени имеет распределение Пуассона, а *время ожидания между событиями* — экспоненциальное распределение. Этот факт проиллюстрирован на рис. 7.5.5. Распределение времени ожидания от некоторого фиксированного момента времени до наступления события подчиняется экспоненциальному закону.

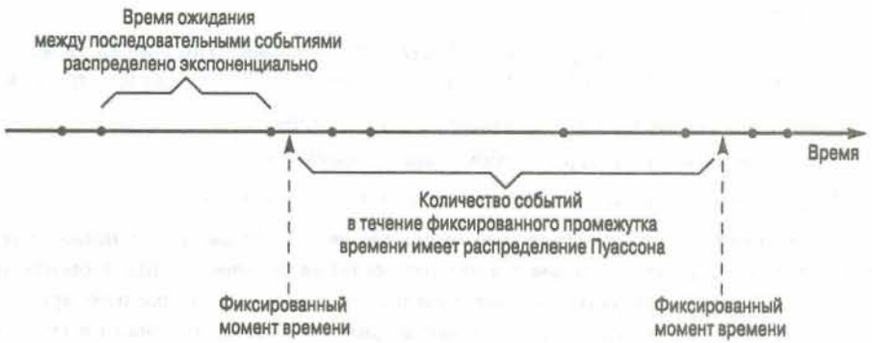


Рис. 7.5.5. Взаимосвязь экспоненциального распределения и распределения Пуассона, когда события происходят независимо и с постоянной частотой

Пример. *Визиты клиентов*

Предположим, что клиенты приходят независимо друг от друга с постоянной средней частотой 40 посетителей в час. Чтобы найти вероятность того, что в течение следующих пяти минут придет по меньшей мере один клиент, необходимо вычислить вероятность того, что подчиняющееся экспоненциальному распределению время ожидания следующего клиента будет меньше пяти минут. Поскольку в течение каждого часа в среднем приходят 40 клиентов, среднее значение интересующей нас экспоненциально распределенной случайной величины составляет 1/40 = 0,025 часа, или 0,025 × 60 = 1,5 минуты. Таким образом, искомая вероятность равна $P(X<5)=1-e^{-5/1,5}=0,96$. Это значение можно также вычислить с помощью Excel, воспользовавшись формулой `=1-EXP(-5/1,5)`. Таким образом, вероятность появления в течение следующих пяти минут по меньшей мере одного клиента оказывается довольно высокой (96%).

7.6. Дополнительный материал

Резюме

Случайная величина — это характеристика, или описание, численного результата случайного эксперимента. Конкретное значение, принимаемое случайной величиной, называется **результатом наблюдения**. Структура вероятностей различных значений случайной величины называется **распределением вероятностей**. Случайная величина может быть **дискретной** (если можно перечислить все возможные результаты) или **непрерывной** (если результатом может быть любое число из некоторого интервала). Некоторые случайные величины в действительности дискретны, но с ними можно и удобнее работать как с непрерывными.

Распределение вероятностей для дискретной случайной величины представляет собой перечень возможных значений с указанием вероятностей их наблюдения. Среднее, или ожидаемое, значение и стандартное отклонение для случайной величины вычисляется следующим образом:

$$\mu = E(X) = \text{Сумма (значение, умноженное на вероятность)} = \sum XP(X);$$

$$\sigma = \sqrt{\text{Сумма (квадрат отклонения, умноженный на вероятность)}} =$$

$$= \sqrt{\sum (X-\mu)^2 P(X)}.$$

Интерпретация этих величин достаточно проста. **Среднее значение**, или **математическое ожидание**, определяет типичное, или среднее, значение, а **стандартное отклонение** задает риск в смысле того, насколько далеким от среднего может приблизительно оказаться результат конкретного наблюдения.

Случайная величина X имеет **биномиальное распределение**, если она представляет *количество наступлений* некоторого события в n попытках, при условии, что (1) в каждой из n попыток существует одинаковая вероятность π наступления события и (2) попытки независимы друг от друга. **Биномиальная доля** равна p = X/n, ее можно выражать и в процентах. Среднее значение и стандартное отклонение имеющей биномиальное распределение случайной величины или биномиальной доли можно найти по следующим формулам.

Среднее значение и стандартное отклонение для биномиального распределения

	Количество наступлений события, X	Доля или процент, $p = X/n$
Среднее	$E(X) = \mu_x = n\pi$	$E(p) = \mu_p = \pi$
Стандартное отклонение	$\sigma_x = \sqrt{n\pi(1-\pi)}$	$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

Формула для расчета вероятности того, что имеющая биномиальное распределение случайная величина X равна некоторому числу a (от 0 до n), приведена ниже.

Вероятность того, что биномиально распределенная величина X равна a:

$$P(X=a) = \binom{n}{a}\pi^a(1-\pi)^{n-a}$$

$$= \frac{n!}{a!(n-a)!}\pi^a(1-\pi)^{n-a}.$$

$$= \frac{1\times 2\times 3\times \ldots \times n}{[1\times 2\times 3\times \ldots \times a][1\times 2\times 3\times \ldots \times (n-a)]}\pi^a(1-\pi)^{n-a}$$

Запись $n!$ (читается как "n факториал") означает произведение целых чисел от 1 до n. По определению 0! = 1. Величина

$$\binom{n}{a} = \frac{n!}{a!(n-a)!}$$

называется биномиальным коэффициентом и читается как "выбор из n по a".

Нормальное распределение — это непрерывное распределение, представленное колоколообразной кривой. Вероятность того, что имеющая нормальное распределение случайная величина примет значение, попадающее в интервал между двумя определенными значениями, равна площади под графиком нормального распределения между этими двумя точками. Для каждой пары чисел, определяющей среднее значение μ и (положительное) стандартное отклонение σ, существует свое нормальное распределение. **Стандартное нормальное распределение** — это нормальное распределение со средним значением $\mu = 0$ и стандартным отклонением $\sigma = 1$. Стандартное нормальное распределение представляет случайную величину, равную отклонению от среднего, измеренному в стандартных отклонениях. **Таблица вероятностей для стандартного нормального распределения** позволяет найти вероятность того, что имеющая стандартное нормальное распределение случайная величина X примет значение, *меньшее* заданного числа z.

Для решения словесно сформулированных задач, связанных с нормальным распределением, прежде всего следует найти среднее значение μ и стандартное отклонение σ, а также выяснить, какая вероятность представляет интерес. Далее необходимо перейти к **нормированному значению** z (количество стандартных отклонений выше среднего или ниже среднего, если нормированное значение отрицательно). Для этого из значения вычитают среднее и делят полученный результат на стандартное отклонение:

$$z = \text{Нормированное значение} = \frac{\text{Значение} - \text{Среднее}}{\text{Стандартное отклонение}} = \frac{\text{Значение} - \mu}{\sigma}.$$

Используя это нормированное значение (или значения), находят соответствующие строки в таблице стандартного нормального распределения и затем получают ответы для указанных в представленной ниже таблице различных типов задач (z, z_1 и z_2 обозначают нормированные значения).

Вычисление вероятностей для нормального распределения

Чтобы найти вероятность того, что Z...	Необходимо...
меньше z	найти для z соответствующую вероятность в таблице
больше z	вычесть предыдущий результат из 1
лежит между z_1 и z_2	найти в таблице вероятности, соответствующие z_1 и z_2, и вычесть меньшую вероятность из большей
лежит за пределами интервала от z_1 до z_2	вычесть предыдущий результат (для вопроса "между z_1 и z_2") из 1

При больших n и значениях π, не слишком близких к 0 или 1, биномиальное распределение можно аппроксимировать нормальным распределением с такими же значениями среднего и стандартного отклонения. Поскольку нормальное распределение — непрерывное, необходимо расширить пределы интервала на S в каждом направлении. Так, например, вероятность того, что имеющая биномиальное распределение величина в точности равна целому числу a, аппроксимируется вероятностью того, что соответствующая нормально распределенная величина лежит в пределах от a −0,5 до a + 0,5.

Если события происходят случайно, независимо и с постоянной средней частотой, количество наблюдений события в течение фиксированного промежутка времени подчиняется **распределению Пуассона**. Это — дискретная случайная величина. Ее стандартное отклонение равно квадратному корню из среднего значения. Если среднее значение велико, распределение Пуассона аппроксимируется нормальным распределением, что дает возможность использования таблицы вероятностей для стандартного нормального распределения. Точное значение вероятности для распределения Пуассона вычисляется по следующей формуле:

$$P(X = a) = e^{-\mu}\left(\frac{\mu^a}{a!}\right).$$

Экспоненциальное распределение — это сильно асимметричное распределение непрерывной величины, которое полезно использовать для исследования таких переменных, как, например, время ожидания или длительность телефонного разговора. Это распределение не обладает "памятью" в том смысле, что после ожидания события в течение некоторого промежутка времени без наступления этого события среднее время ожидания следующего события не уменьшается по сравнению с существовавшим на момент начала ожидания. Стандартное отклонение при таком распределении всегда равно среднему значению. Вероятность того, что имеющая экспоненциальное распределение случайная величина X со средним значением μ принимает значение, меньшее a, выражается формулой $P(X < a) = 1 - e^{-a/\mu}$. Экспоненциальное распределение нельзя аппроксимировать нормальным.

Основные термины

- Случайная величина (random variable), 279
- Наблюдаемое значение (observation), 279
- Распределение вероятности (probability distribution), 279
- Дискретная случайная величина (discrete random variable), 279
- Непрерывная случайная величина (continuous random variable), 279
- Среднее значение, или математическое ожидание (mean or expected value), 281
- Стандартное отклонение (standard deviation), 281
- Биномиальное распределение (binomial distribution), 285
- Биномиальная пропорция (binomial proportion), 286
- Нормальное распределение (normal distribution), 295
- Стандартное нормальное распределение (standard normal distribution), 298
- Таблица вероятностей для стандартного нормального распределения (standard normal probability tree), 298
- Нормированное значение (standardized number), 303
- Распределение Пуассона (Poisson distribution), 315
- Экспоненциальное распределение (exponential distribution), 318

Контрольные вопросы

1. а) Что такое случайная величина?

   б) Чем случайная величина отличается от числа?

2. а) Что представляет собой дискретная случайная величина?

   б) Что представляет собой непрерывная случайная величина?

   в) Приведите пример такой дискретной случайной величины, которую для практических целей можно считать непрерывной.

3. а) Что такое распределение вероятности для дискретной случайной величины?

   б) Как найти среднее значение дискретной случайной величины? Как интерпретировать этот результат?

   в) Как найти стандартное отклонение дискретной случайной величины? Как интерпретировать этот результат?

4. а) Как выяснить, имеет ли случайная величина биномиальное распределение?

   б) Что такое биномиальная доля?

   в) Что такое n, π, X и p?

5. Для биномиального распределения

   а) Почему для нахождения величин вероятностей нельзя просто построить дерево вероятностей?

   б) Как найти среднее значение и стандартное отклонение?

   в) Как найти вероятность того, что X равно некоторому числу?

   г) Как найти точную вероятность того, что X больше или равно некоторому числу?

   д) Как при большом n найти приближенную вероятность того, что X больше или равно некоторому числу?

6. а) Что такое факториал?

   б) Найдите 3!, 0! и 15!.

   в) Что такое биномиальный коэффициент? Что он показывает в формуле вероятности биномиального распределения?

   г) Найдите биномиальный коэффициент для "выбора из 8 по 5".

7. а) Что такое нормальное распределение?

   б) Охарактеризуйте все возможные нормальные распределения.

   в) Что показывает площадь под кривой нормального распределения?

   г) Что такое стандартное нормальное распределение? Для чего оно используется?

   д) Какие числа содержит таблица вероятностей для стандартного нормального распределения?

   е) Найдите вероятность того, что имеющая стандартное нормальное распределение случайная величина принимает значение, меньшее чем $-1{,}65$.

   ж) Как нормируется значение?

8. а) В каких случаях появляются величины, имеющие распределение Пуассона?

   б) Является ли распределение Пуассона дискретным или непрерывным?

   в) Чему равно стандартное отклонение в распределении Пуассона?

   г) Как можно вычислить вероятность для имеющей распределение Пуассона величины при большом среднем значении?

   д) Как найти точное значение вероятности для имеющей распределения Пуассона случайной величины?

9. а) В каких случаях появляются случайные величины, имеющие экспоненциальное распределение?

   б) Как вы понимаете утверждение о том, что характеризующаяся экспоненциальным распределением случайная величина "не имеет памяти"?

   в) Можно ли использовать для поиска вероятностей таблицу вероятностей стандартного нормального распределения в случае экспоненциального распределения? Почему да или почему нет?

   г) Как найти вероятность для экспоненциального распределения?

Задачи

1. Оценивается опцион "колл" на покупку обыкновенных акций. Если курс акций понизится, действие опциона окончится без получения дохода. Если курс акций вырастет, выплаты будут зависеть от того, насколько сильно вырастут акции. Для простоты будем считать, что выплаты представляются дискретным распределением, с распределением вероятностей, показанным в табл. 7.6.1. Даже несмотря на то, что цены на торгуемые на бирже опционы ведут себя скорее подобно непрерывным случайным переменным, такое дискретное приближение все равно даст полезные приближенные результаты. Дайте, исходя из представленного в таблице дискретного распределения вероятностей, ответы на приведенные ниже вопросы.

   а) Найдите среднее, или ожидаемое, значение платежа по опциону.

   б) Кратко опишите, что характеризует это ожидаемое значение.

   в) Найдите стандартное отклонение платежа по опциону.

   г) Кратко опишите, что характеризует это стандартное отклонение.

   д) Найдите вероятность того, что платеж по опциону составит по меньшей мере \$20.

   е) Найдите вероятность того, что платеж по опциону составит менее \$30.

Таблица 7.6.1. Распределение вероятностей платежей

Платеж, дол.	Вероятность
0	0,50
10	0,25
20	0,15
30	0,10

2. Время, в течение которого система находится по причине неисправности в неработающем состоянии, приблизительно описывается приведенным в табл. 7.6.2 распределением вероятностей. Предположим, что это точное описание продолжительности простоя системы. Это значит, что существуют три типа легко различаемых проблем, для решения которых требуется именно столько времени (5, 30 или 120 минут).

   а) Какого типа распределение вероятности представлено в этой таблице?

   б) Найдите среднее значение времени простоя по причине неисправности.

   в) Найдите стандартное отклонение времени простоя.

   г) Исходя из приведенных в таблице данных, чему равна вероятность того, что время простоя превысит 10 минут?

   д) Чему равна вероятность того, что время простоя будет отличаться от среднего значения не более чем на одну величину стандартного отклонения? Соответствует ли полученный результат тому, чего следовало бы ожидать для нормального распределения?

3. Вложение денег приведет к выплате \$105 с вероятностью 0,7 или к выплате \$125 с вероятностью 0,3. Найдите риск (выраженный стандартным отклонением) этого вложения.

4. Предположим, что для некоторого дня вероятность отсутствия заказов равна 30%, вероятность поступления одного заказа — 50%, вероятность поступления двух заказов — 15%, а вероятность поступления трех заказов — 5%. Найдите ожидаемое число заказов и изменчивость их количества.

5. Представьте себе, что вы работаете в отделе кредитов крупного банка. Известно, что один из заемщиков испытывает финансовые затруднения и, возможно, не сможет произвести текущий платеж, срок которого наступает на следующей неделе. Вы считаете, что с вероятностью 60% он внесет всю подлежащую выплате сумму в \$50 000, с вероятностью 30% внесет только половину, а с вероятностью 10% не внесет ничего.

   а) Найдите ожидаемую кредитную выплату.

   б) Найдите уровень риска для данной ситуации.

6. Ожидаемая доходность планируемых инвестиций в новую работающую в области высоких технологий компанию показана в табл. 7.6.3 (100% означает удвоение внесенной суммы, −50% означает потерю половины инвестированных денег).

   а) Найдите среднюю доходность и дайте пояснения к найденному значению.

Таблица 7.6.2. Распределение вероятности для времени простоя

Неисправность	Время простоя, мин.	Вероятность
Незначительная	5	0,60
Существенная	30	0,30
Катастрофическая	120	0,10

Таблица 7.6.3

Доходность, %	Вероятность
100	0,20
50	0,40
0	0,25
−50	0,15

   б) Найдите стандартное отклонение доходности и дайте пояснения к найденному значению.

   в) Руководствуясь таблицей, найдите вероятность того, что доход превысит 40%.

   г) Как оценить риск такого вложения?

7. Существует возможность вложения средств в один из четырех проектов, связанных с использованием находящегося в вашей собственности земельного участка. Для простоты выплаты по проектам (как чистая сумма в долларах на сегодняшний день) представлены в виде дискретного распределения. Если землю продать, можно гарантированно получить \$60 000. Если построить многоквартирный дом, выплата оценивается в \$130 000 в случае удачи (с вероятностью 0,60) и \$70 000 в противном случае. Если построить дом на одну семью, выплаты составят \$100 000 (с вероятностью 0,60) или \$60 000 в противном случае. Наконец, можно построить казино, которое будет давать хороший доход — \$500 000 — однако с вероятностью только 0,10, поскольку государственные органы достаточно редко дают разрешение на работу казино. В случае же отсутствия разрешения деньги просто будут потеряны.

   а) Найдите ожидаемые платежи для каждого из четырех проектов. Установите приоритетность проектов, исходя только из ожидаемых платежей.

   б) Найдите стандартное отклонение для каждого из этих четырех проектов. Установите приоритетность проектов, исходя только из их рискованности.

   в) Можно ли какой-либо проект (или проекты) полностью исключить из рассмотрения, если учитывать и ожидаемые платежи, и рискованность?

   г) Что можно сказать об остальных проектах? В частности, можно ли указать один проект как лучший во всех отношениях?

8. Руководитель отдела контроля качества выявил четыре основные проблемы, возникающие при выпуске устройств, их распространенность (т.е. вероятность того, что данная проблема имеет место для одного произведенного устройства) и затраты, необходимые для устранения каждой из проблем (табл. 7.6.4). Предположим, что в каждом случае может возникнуть только одна проблема.

   а) Вычислите стоимость устранения для каждой отдельной проблемы. Так, например, ожидаемые затраты на устранение проблемы поломки корпуса

Таблица 7.6.4. Проблемы качества продукции: тип, распространенность и затраты

Неисправность	Вероятность	Стоимость устранения, дол.
Поломка корпуса	0,04	6,88
Отказ электроники	0,02	12,30
Отсутствие соединения	0,06	0,75
Физический дефект	0,01	2,92

составляют $0{,}04 \times 6{,}88$. Сравните результаты и укажите самую серьезную из проблем с точки зрения ожидаемых затрат на устранение.

б) Найдите общие ожидаемые затраты на устранение всех четырех проблем.

в) Найдите стандартное отклонение затрат на устранение проблем (не забудьте учесть устройства, не требующие устранения проблем).

г) Составьте краткий отчет для высшего руководства с описанием и анализом ситуации.

9. Предположим, что 8% кредитов, которые вы одобряете в качестве вице-президента отдела кредитов регионального банка, не будут возвращены никогда. Предположим также, что в прошлом году было одобрено предоставление 284 кредитов и что возвраты денег по всем кредитам производятся независимо друг от друга.

   а) Какое количество одобренных вами кредитов не будет возвращено? Какой процент невыплат можно ожидать?

   б) Найдите обычно используемую меру уровня неопределенности для количества таких одобренных кредитов, которые не будут возвращены. Дайте краткую интерпретацию полученного значения.

   в) Найдите обычно используемую меру уровня неопределенности для процента таких одобренных кредитов, которые не будут возвращены. Дайте краткую интерпретацию полученного значения.

10. В ходе предстоящих на следующей неделе выборов предвидится разделение голосов "за" и "против" почти поровну. Предположим, что 50% избирателей отдадут свои голоса "за", а 50% — "против". Вы проводите социологический опрос, в котором изъявили желание принять участие 791 случайно выбранных избирателей. Насколько, приблизительно, будет отличаться процент ответивших в ходе опроса, что будет голосовать "за", от значения 50% в генеральной совокупности, которое вы стараетесь оценить?

11. Решите еще раз предыдущую задачу, предположив на этот раз, что среди всех избирателей "за" проголосует 85%. Уменьшилась или увеличилась неопределенность по сравнению с тем случаем, когда предполагалось 50%? Почему?

12. Вы провели социологический опрос 358 выбранных случайным образом человек. В ходе опроса установлено, что 94 из них заинтересованы в возможности получения новой услуги кабельного телевидения. Какова неопределенность этого значения "94" в сравнении со средним значением, которое следует ожидать в таком исследовании? (Можно предположить, что из всех тех людей, которых вы могли бы опросить, интерес проявили бы ровно 25%).

13. На сегодня планируется позвонить в восемь фирм с предложением продать им некоторую продукцию. В качестве грубого приближения принимается, что каждый такой звонок с вероятность 20% приводит к продаже. Также считается, что фирмы принимают решения о покупке независимо друг от друга. Найдите вероятность того, что день окажется ужасным и ни одной продажи не будет.

14. Сегодня на фондовой бирже неудачный день, стоимость 80% ценных бумаг падает. Вы оцениваете портфель, содержащий 15 ценных бумаг. Предполагается биномиальное распределение количества понижающихся в цене бумаг.

   а) Какие допущения делаются, если в данном случае используют биномиальное распределение?

   б) Для какого количества ценных бумаг, входящих в портфель, ожидается снижение стоимости?

   в) Чему равно стандартное отклонение количества таких входящих в портфель ценных бумаг, стоимость которых снижается?

   г) Найдите вероятность падения в цене всех 15 ценных бумаг.

   д) Найдите вероятность падения в цене точно 10 ценных бумаг.

   е) Найдите вероятность падения в цене 13 или более ценных бумаг.

15. Ваша фирма приняла решение о проведении выборочного опроса 10 потребителей для выяснения вопроса о том, следует ли вносить изменения в некоторый потребительский товар. Основной конкурент фирмы недавно провел аналогичное исследование в несколько большем объеме, в результате которого было установлено, что 86% потребителей изменение одобряют. Однако ваша фирма не имеет доступа к этой информации (но вы можете воспользоваться этим значением для решения данной задачи).

   а) Какое распределение будет иметь в вашем исследовании количество потребителей, одобряющих изменение?

   б) Чему равно ожидаемое *количество* людей не из числа 10 опрошенных, которые одобрят изменение?

   в) Чему равно стандартное отклонение количества людей не из числа 10 опрошенных, которые одобрят изменение?

   г) Чему равен ожидаемый *процент* людей не из числа 10 опрошенных, которые одобрят изменение?

   д) Чему равно стандартное отклонение для процента людей не из числа 10 опрошенных, которые одобрят изменение?

   е) Чему равна вероятность того, что точно восемь опрошенных потребителей одобрят изменения?

   ж) Чему равна вероятность того, что изменение одобрят восемь или более опрошенных потребителей?

16. В обычных условиях нефтеперегонный завод может перерабатывать в среднем 135 000 баррелей сырой нефти в день со стандартным отклонением 6000 баррелей в день. Можно предположить, что объем обрабатываемой нефти подчиняется нормальному распределению.

а) Найдите вероятность того, что за один день будет переработано более чем 135000 баррелей нефти.

б) Найдите вероятность того, что за один день будет переработано более чем 130000 баррелей нефти.

в) Найдите вероятность того, что за один день будет переработано более чем 150000 баррелей нефти.

г) Найдите вероятность того, что за один день будет переработано менее чем 125000 баррелей нефти.

д) Найдите вероятность того, что за один день будет переработано менее чем 100000 баррелей нефти.

17. Отдел контроля качества закупаемых по контракту клапанов требует, чтобы диаметр клапана находился в пределах от 2,53 до 2,57 сантиметра. Предположим, что оборудование, на котором производятся клапаны, налажено таким образом, что средний диаметр клапана равен 2,56 сантиметра, а стандартное отклонение составляет 0,01 сантиметра. В предположении о нормальном распределении необходимо определить, какой процент произведенных в течение длительного времени клапанов будет соответствовать требованиям?

18. Предположим, что сегодня в момент закрытия биржи значение индекса активности было равно 9246 пунктам. Завтра ожидается подъем в среднем на 4 пункта со стандартным отклонением 115 пунктов (предполагается, что распределение нормальное).

а) Найдите вероятность завтрашнего понижения фондового рынка.

б) Найдите вероятность того, что завтра на фондовом рынке будет наблюдаться повышение более чем на 50 пунктов.

в) Найдите вероятность того, что завтра на фондовом рынке будет наблюдаться повышение более чем на 100 пунктов.

г) Найдите вероятность того, что завтра на фондовом рынке будет наблюдаться понижение более чем на 150 пунктов.

д) Найдите вероятность того, что завтра на фондовом рынке будет наблюдаться колебание, превышающее 200 пунктов в любую сторону.

19. Из предыдущего опыта в следующую субботу ожидаются поступления, составляющие в среднем $2353,25, со стандартным отклонением $291,63. Распределение нормальное.

а) Найдите вероятность того, что это будет обычная суббота, что определяется поступлениями в пределах от $2000 до $2500.

б) Найдите вероятность того, что это будет совершенно потрясающая суббота в связи с ожидаемыми поступлениями в размере свыше $2500.

в) Найдите вероятность того, что это будет весьма посредственная суббота с поступлениями менее $2000.

20. Количество металла (в тоннах) в некоторой части прииска, как предполагается, подчиняется нормальному распределению со средним значением 185 и стандартным отклонением 40. Найдите вероятность того, что количество металла менее чем 175 тонн.

21. Вы — фермер и собираетесь убирать урожай. Для того чтобы учесть неопределенность размера урожая, вы предполагаете, что он может подчиняться нормальному распределению со средним значением 80 000 бушелей и стандартным отклонением 2500 бушелей. Найдите вероятность того, что урожай превысит 84 000 бушелей.

22. Предположим, что рабочая частота электронных микросхем имеет нормальное распределение со средним значением 450 мегагерц и стандартным отклонением 9 мегагерц. Какой процент вашей продукции окажется "супермикросхемами" с рабочей частотой 466 мегагерц или более?

23. Точный объем платежей, которые поступят в следующем месяце, неизвестен, однако из предыдущего опыта следует, что он составит примерно на $2500 больше или меньше чем $13000 и будет иметь нормальное распределение. Найдите вероятность того, что в следующем месяце будет получено от $10000 до $15000.

24. Новый проект будет считаться "успешным", если в течение двух лет будет захвачено 10% или больше рынка. Отдел маркетинга рассмотрел все возможности и пришел к выводу о том, что для данного продукта за это время можно ожидать захвата 12% рынка. Однако это не точное значение. Прогнозируется также, что стандартное отклонение составит 3%. Таким образом, неопределенность прогноза в 12% составляет 3 пункта. Распределение считаем нормальным.

а) Найдите вероятность того, что новый проект окажется успешным.

б) Найдите вероятность того, что новый проект не будет иметь успеха.

в) Найдите вероятность того, что новый проект будет иметь сумасшедший успех, что соответствует захвату по меньшей мере 15% рынка.

г) Для точности маркетинговой оценки найдите вероятность того, что достигнутая доля рынка будет близка к прогнозируемому значению 12%, т.е. попадет в интервал от 11% до 13%.

25. В процессе производства выпускаются полупроводниковые микросхемы. При этом 6,3% микросхем имеют дефекты. Предположим, что дефекты появляются в разных микросхемах независимо. Завтра предстоит выпустить 2000 микросхем.

а) Какое распределение имеет количество дефектных микросхем в завтрашней партии?

б) Сколько ожидается выпустить дефектных микросхем?

в) Найдите стандартное отклонение количества дефектных микросхем.

г) Найдите (приблизительно) вероятность того, что в выпущенной завтра партии будет менее 130 дефектных микросхем.

д) Найдите (приблизительно) вероятность того, что в выпущенной завтра партии будет более 120 дефектных микросхем.

е) Только что сообщили, что из завтрашней партии в 2000 микросхем потребуется отгрузить заказчику 1860 работающих. Какова вероятность успешного выполнения заказа? Есть ли необходимость увеличить запланированный на завтра объем производства?

ж) Чему будет равна вероятность отгрузки 1860 работающих микросхем, если в производственный план на завтра включить 2100 микросхем?

26. На завтра запланировано проведение голосования по вопросу о профсоюзной забастовке. Похоже на то, что она может состояться. Предположим, что количество голосов, поданных за проведение забастовки, следует биномиальному распределению. Ожидается, что в голосовании примут участие 300 человек, а вероятность того, что каждый из них будет голосовать "за", составляет 0,53.

а) Найдите n и π для этой биномиально распределенной случайной величины.

б) Найдите среднее и стандартное отклонение для количества тех, кто выскажется за проведение забастовки.

в) Найдите (приблизительно) вероятность того, что забастовка будет проведена (т.е. что большинство участников голосования выскажутся за ее проведение).

27. Снова вернемся к предыдущей задаче и найдем ответы на поставленные вопросы в предположении, что в голосовании примут участие 1000 человек. (Вероятность для каждого отдельного участника голосования остается прежней.)

28. Предположим, что в случае опроса всего населения Детройта 18,6% выразили бы готовность покупать ваш товар. Планируется проведение выборочного опроса 250 человек. Найдите (приближенно) вероятность того, что полученное в результате выборочного опроса значение будет излишне оптимистичным, что определяется как готовность к покупке товара более чем 22,5% опрошенных.

29. Предположим, что 15% хранящегося на большом складе товара имеет дефекты. Вы выбрали случайным образом 250 единиц товара для проведения детального исследования. Найдите (приблизительно) вероятность того, что дефекты имеются более чем в 20% образцов, попавших в данную выборку.

30. Планируется проведение опроса среди 350 потребителей, выбранных случайным образом из списка возможных клиентов. Цель опроса состоит в оценке этого списка и принятии решения о том, имеет ли смысл поручать агентам по продаже связываться со всеми, кто в него включен. В предположении, что 13% людей из всего списка дадут положительный ответ, вычислите (приблизительно) вероятности перечисленных ниже событий.

а) Более 10% выбранных случайным образом потребителей дадут положительный ответ.

б) Более 13% выбранных случайным образом потребителей дадут положительный ответ.

в) Более 15% выбранных случайным образом потребителей дадут положительный ответ.

г) От 10% до 15% выбранных случайным образом потребителей дадут положительный ответ.

31. Проведена пробная рассылка по почте каталога в адрес 1000 человек, выбранных случайным образом из базы данных, содержащей 12320 адресов. В случае поступления в течение двух недель заказов от 2,7% или более человек, включенных в пробную рассылку, планируется рассылка каталога по остальным 11320 адресам. Найдите (приблизительно) вероятность того, что массовая рассылка будет проведена в случае каждого из перечисленных ниже сценариев.

а) Предположим, что в действительности в течение двух недель заказ поступил бы от 2% генеральной совокупности (т.е. всех включенных в базу данных людей).

б) Предположим, что в действительности в течение двух недель заказ поступил бы от 3% генеральной совокупности.

в) Предположим, что в действительности в течение двух недель заказ поступил бы от 4% генеральной совокупности.

32. В следующем месяце ожидается в среднем 1671 обращение по вопросу гарантийного ремонта, и истинное значение количества обращений имеет распределение Пуассона.

а) Найдите стандартное отклонение для количества таких обращений.

б) Найдите (приблизительно) вероятность того, что количество обращений по вопросу гарантийного ремонта превысит 1700.

33. Если завтрашний день будет обычным, в отдел работы с персоналом ожидается поступление по почте резюме от 175 претендентов на вакантные должности. Можно предположить, что все претенденты действуют независимо друг от друга.

а) Каким распределением вероятности описывается количество резюме, поступающих в отдел работы с персоналом?

б) Чему равно стандартное отклонение количества получаемых резюме?

в) Найдите (приблизительно) вероятность того, что день будет спокойным и в отдел поступит только 160 или менее резюме.

34. В обычный день магазин по продаже одежды обслуживает в среднем 2,6 "особых клиентов". Этих клиентов сразу отводят в специальную комнату, предлагают им специальное обслуживание, подают чай (или кофе) с печеньем, в эту же комнату приносят и одежду. Можно предположить, что количество таких покупателей, которые посетят магазин завтра, имеет распределение Пуассона.

а) Найдите стандартное отклонение количества таких особых клиентов.

б) Найдите вероятность того, что завтра таких посетителей не будет.

в) Найдите вероятность того, что завтра магазин посетят ровно 4 особых клиента.

35. Для получения достаточного дохода, который позволит выплатить в этом году долги фирмы, необходимо заключить по меньшей мере два контракта. Обычно это не составляет больших проблем, поскольку в среднем фирма заключает 5,1 контракта в год. Можно предположить, что количество контрактов описывается распределением Пуассона.

    а) Найдите вероятность того, что доход будет достаточным для оплаты долговых обязательств фирмы в этом году.

    б) Найдите вероятность заключения трех контрактов.

36. Длительность промежутков времени между появлениями клиентов имеет экспоненциальное распределение. На данный момент среднее время между появлениями клиентов составляет 6,34 минуты.

    а) С момента появления последнего клиента прошло три минуты. Найдите среднее время до появления следующего клиента.

    б) С момента появления последнего клиента прошло десять минут. Найдите среднее время до появления следующего клиента

37. В описанном в предыдущей задаче случае клиент появился только что.

    а) Найдите вероятность того, что промежуток времени до появления следующего клиента окажется меньше трех минут.

    б) Найдите вероятность того, что промежуток времени до появления следующего клиента окажется больше десяти минут.

    в) Найдите вероятность того, что промежуток времени до появления следующего клиента составит от пяти до шести минут.

38. В соответствии со спецификациями срок службы кинескопа телевизора составляет 50000 часов. Предположим, что время до выхода кинескопа из строя подчиняется экспоненциальному распределению.

    а) Симметрично ли это распределение?

    б) Чему равно стандартное отклонение срока продолжительности службы кинескопа?

    в) Кинескоп работал непрерывно в течение 8500 и все еще находится в рабочем состоянии. Чему равно ожидаемое время работы кинескопа с настоящего момента до выхода из строя? (При ответе на этот вопрос будьте очень внимательны.)

39. Для описанной в предыдущей задаче ситуации, в предположении соответствующего данному случаю распределения, дайте ответы на следующие вопросы.

    а) Чему равна вероятность работы кинескопа в течение 100000 часов и более? (Этот промежуток времени превышает номинальное значение срока службы в два раза.)

    б) Гарантия на кинескоп дается на 5000 часов работы. Какой процент кинескопов, как ожидается, выходит из строя в течение гарантийного периода?

40. Сравните "вероятность того, что результат находится в пределах одной величины стандартного отклонения от среднего значения" для экспоненциального и для нормального распределения.

Упражнения с использованием базы данных

Обратитесь к базе данных наемных работников, приведенной в приложении А.

1. Будем считать данные, приведенные в каждом из столбцов, набором независимых результатов наблюдения случайной величины.

    а) Какая величина дана в каждом из случаев — непрерывная или дискретная? Почему вы так считаете?

    б) Рассмотрим событие "годовая зарплата более 40000 дол." Найдите значение имеющей биномиальное распределение случайной величины X, равной количеству наступлений данного события. Найдите также биномиальную долю p и поясните, что она показывает.

    в) Какую часть работников составляют мужчины? Рассмотрите эту величину как биномиальную долю. Чему равно n?

2. У вас есть вакансия, и вы хотите принять на работу нового сотрудника. Предположите, что опыт работы новых сотрудников имеет нормальное распределение со средним и стандартным отклонением (выборочным), как у ваших нынешних сотрудников.

    а) Найдите вероятность того, что опыт работы нового сотрудника будет превышать шесть лет.

    б) Найдите вероятность того, что опыт работы нового сотрудника будет менее трех лет.

    в) Найдите вероятность того, что опыт работы нового сотрудника будет составлять от четырех до семи лет.

3. Предположим, что новый сотрудник может с одинаковой вероятностью оказаться как мужчиной, так и женщиной, а количество представителей каждого пола следует биномиальному распределению. (Не забывайте, что база данных содержит *результаты наблюдений* случайных величин, а не сами случайные величины.)

    а) Найдите n и π для биномиального распределения количества мужчин.

    б) Найдите n и π для биномиального распределения количества женщин.

    в) Найдите наблюдаемое значение количества женщин X.

    г) Воспользуйтесь для биномиального распределения нормальным приближением и найдите вероятность того, что база данных содержит именно такое (ответ на вопрос "в") количество женщин или меньшее.

Проекты

1. Выберите непрерывную случайную величину, с которой вы можете иметь дело в своей нынешней или будущей работе в качестве руководителя. Рассматривайте ее как имеющую нормальное распределение случайную величину и оцените (т.е. сделайте соответствующее предположение) среднее значение и стандартное отклонение. Определите три представляющих интерес события, связанных с этой случайной величиной, и вычислите вероятности этих событий. Кратко опишите, что вам удалось выяснить.

2. Выберите дискретную случайную величину (принимающую от 3 до 10 различных возможных значений), с которой вы можете иметь дело в своей нынешней или будущей работе в качестве руководителя. Оцените (т.е. сделайте соответствующее предположение), какому распределению она подчиняется. Вычислите среднее значение и стандартное отклонение. Определите два представляющих интерес события, связанных с этой случайной величиной, и вычислите вероятности событий. Кратко опишите, что вам удалось выяснить.
3. Выберите имеющую биномиальное распределение случайную величину, с которой вы можете иметь дело в своей нынешней или будущей работе в качестве руководителя. Оцените (т.е. сделайте соответствующее предположение) величины n и π. Вычислите среднее значение и стандартное отклонение. Определите два представляющих интерес события, связанных с этой случайной величиной, и вычислите вероятности этих событий. Кратко опишите, что вам удалось выяснить.
4. Найдите в Internet и запишите результаты наблюдения по меньшей мере пяти различных случайных величин, таких, как биржевые котировки, процентные ставки, объемы продаж корпораций, или любых других величин, связанных с коммерческой деятельностью и представляющих для вас интерес.

Ситуация для анализа

Стоимость опциона на аренду нефтяного месторождения

Существует возможность аренды нефтяного месторождения, которая кажется настолько заманчивой, что это мало похоже на правду. Судите сами: нефтяное месторождение, запасы которого оцениваются в 1 500 000 баррелей, можно взять в аренду на 3 года всего лишь за $1 000 000. Возможность представляется очень выгодной: платим миллион, добываем нефть, продаем ее по текущей цене "спот", составляющей $18,36 за баррель, — и можно отдыхать.

Однако при более внимательном рассмотрении становится понятно, почему никто не ухватился за эту "возможность". Оказывается, что добывать нефть будет непросто в силу геологических особенностей и удаленности месторождения. При тщательном изучении вопроса выясняется, что прогнозируемые затраты на добычу нефти составят $30000000. Можно сделать вывод о том, что разработка этого месторождения окажется *убыточной*. Казалось бы, на этом можно и закрыть вопрос.

На следующей неделе, несмотря на занятость рассмотрением другого инвестиционного проекта, вы все равно снова и снова возвращаетесь мыслями к этому проекту. В частности, низкая стоимость аренды и тот факт, что срок аренды составляет три года, просто вынуждают вас провести анализ ситуации с использованием сценария *Что, если?* и с учетом того, что сейчас обязательств по добыче нефти нет, и, кроме того, организовать добычу нефти можно достаточно быстро (примерно за месяц) в любое время в течение трехлетнего срока аренды. Появляются предположения, что это все же может оказаться выгодным. Что если цены на нефть в течение трех лет вырастут настолько, что разработка этого месторождения окажется выгодной? Если так — можно будет добывать нефть. Однако если цены не вырастут в достаточной степени, срок аренды по окончании трех лет истечет, а нефть останется под землей. Таким образом, будущие цены на нефть определят, следует ли использовать возможность добычи нефти ("опцион на добычу") или нет.

Однако решение о покупке такого опциона — рискованное. Насколько велик риск? Каким может оказаться возможный доход? Для оценки ситуации определяем структуру основных вероятностей для цены на нефть — источника неопределенности в этой ситуации.

Цена на нефть в будущем, дол.	Вероятность
10	0,1
15	0,2
20	0,4
25	0,2
30	0,1

Вопросы для обсуждения

1. Сколько можно было бы заработать при отсутствии затрат на добычу нефти? Хватит ли этого, чтобы "уйти на отдых"?
2. Учитывая затраты на добычу, будут ли финансовые потери, если приобрести право аренды и сразу же начать добывать нефть? Сколько составят эти потери?
3. Продолжим рассмотрение сценария и рассчитаем будущие чистые поступления для каждой из предполагаемых в будущем цен на нефть. Для этого умножим каждый раз цены на нефть на количество баррелей и вычтем из полученного значения стоимость добычи. Если результат отрицательный — разрабатывать месторождение нет смысла, так что отрицательные значения можно переводить в нулевые. (На этом этапе стоимость аренды не вычитается, поскольку ее уже считаем выплаченной.)
4. Теперь найдем средние возможные чистые доходы за вычетом стоимости аренды. Сколько в среднем можно получить (или потерять), взяв в аренду это нефтяное месторождение? (Инфляцию можно не учитывать.)
5. Насколько рискованным оказывается предложение об аренде?
6. Будете ли вы брать в аренду это месторождение?

Статистический вывод

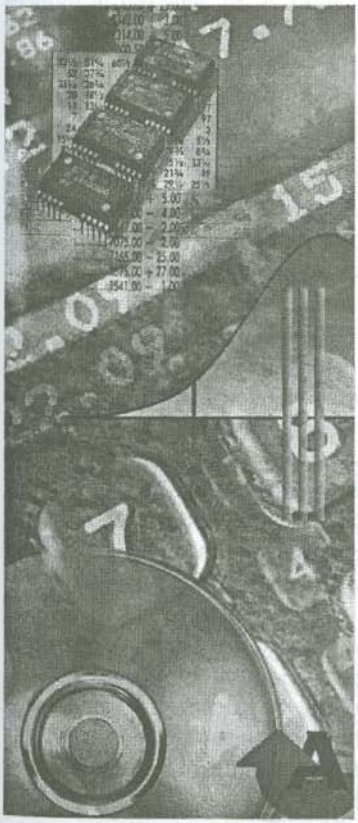
В этой части...

Глава 8. "Построение случайной выборки: предварительное планирование для сбора данных"

Глава 9. "Доверительные интервалы: допущение о неточности оценок"

Глава10. "Проверка статистических гипотез: выбор между реальностью и совпадением"

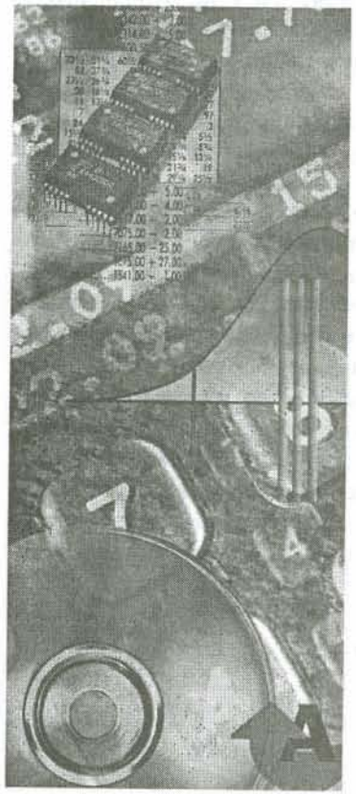
Реальная сила статистики заключается в применении концепции *вероятности* к ситуации, когда есть некоторые *данные*. Результаты, которые называют *статистическим выводом*, дают на основе имеющихся данных строгие вероятностные утверждения о мире. Даже небольшой набор данных обеспечит хорошие результаты в случае правильного их использования. Вот почему, например, исходя из результата политических и маркетинговых опросов тщательно отобранной выборки населения можно сказать, что думают или что делают "все американцы". Одним из наилучших способов отобрать из большой группы небольшую репрезентативную выборку является построение *случайной выборки*, речь о которой пойдет в главе 8. Строгое вероятностное утверждение о неизвестном количестве дает *доверительный интервал*, о котором мы поговорим в главе 9. Если необходимо сделать выбор между двумя возможностями, необходимо использовать процедуру *проверки статистических гипотез* (см. главу 10), которая подскажет, что о ситуации говорят данные. В условиях неопределенности, когда точные ответы отсутствуют, статистический вывод дает ответы с по крайней мере *известным* уровнем ошибки.



ГЛАВА 8

Построение случайной выборки: предварительное планирование для сбора данных

Девять часов утра, среда... Вы только что прочли служебную записку своего начальника, в которой он просит вас отследить реакцию клиентов вашей фирмы на новый график предложенных ценовых скидок. Вам необходимо завтра к 10.00 утра представить доклад на заседание совета директоров. Что предпринять в этой ситуации? Некоторые моменты очевидны. Например, вам необходимо переговорить с некоторыми клиентами. Обдумывая эту задачу, вы решаете, что на опрос по телефону только одного клиента необходимо 10 минут. В вашу базу данных внесено 1687 клиентов, и, конечно, за такое короткое время, располагая небольшим количеством свободных сотрудников, вы не сможете обзвонить всех ваших клиентов. Что же делать?



Есть выход: вы построите *выборку* из *генеральной совокупности* всех клиентов фирмы, внесенных в базу данных. Таким образом, вам нужно будет обзвонить отобранное реальное количество клиентов. При этом вам приходится надеяться, что выборка является достаточно *репрезентативной (представительной)* для более крупной генеральной совокупности и поэтому в вашем докладе на заседании совета директоров необходимые факты будут отражены. Ведь совет интересует реакция *всех* потребителей, а не только тех, которые попали в вашу небольшую выборку. Однако только лишь надежда на удачу не гарантирует того, что выборка является действительно репрезентативной. От способа построения выборки зависит, будет она действительно полезной или нет.

Как построить выборку? Можно составить список потребителей, с которыми необходимо поговорить по другим вопросам, а заодно узнать их отношение к предложенному графику ценовых скидок. Однако вы правильно сделаете, отвергнув эту идею, поскольку такой список не будет репрезентативным, так как он состоит в основном из "скрипучих колес", которые требуют больше поддержки и являются менее экономически независимыми, чем основная часть ваших клиентов. И, что еще хуже: эта группа клиентов стремится размещать более мелкие заказы на низкотехнологичное оборудование по сравнению с типичными потребителями вашей продукции. Чтобы получить действительно репрезентативную группу клиентов, выборку нужно строить иначе.

Случайный отбор поможет вам. Для этого есть две причины. Во-первых, случайный отбор гарантирует репрезентативность выборки (по крайней мере, репрезентативность среднего), поскольку не опирается на какое-либо определенное свойство, которое могло бы привести к *смещению* выборки и запутать результат.[1] Во-вторых, использование тщательно контролируемой статистической процедуры позволит вам приблизительно оценить, *насколько отличаются* ваши результаты (полученные из выборки) от свойств генеральной совокупности. Таким образом, используя метод случайного отбора, вы получите приблизительно корректные результаты (по сравнению с результатами, которые вы получили бы, опросив всех потребителей) и будете знать, достаточно ли они точны для принятия необходимых решений.

В частности, если вы заинтересованы в оценке среднего большой генеральной совокупности путем использования среднего случайной выборки, то это выборочное среднее имеет (приблизительно) *нормальное распределение* с уменьшающейся вариацией (и, следовательно, увеличивающейся точностью) по мере увеличения размера выборки n. Это свойство гарантирует такой математический факт, как *центральная предельная теорема*. Приблизительно оценить, насколько далеко расположено наблюдаемое среднее выборки от неизвестного среднего генеральной совокупности, можно с помощью *стандартной ошибки среднего*, которая играет решающую роль в статистическом выводе, так как измеряет качество вашей информации, отражая, насколько удачна или неудачна ваша оценка.

8.1. Генеральные совокупности и выборки

Генеральная совокупность — это набор объектов (людей, предметов или чего-либо еще), о которых вы хотите получить информацию. **Выборка** — это небольшой набор объектов, извлеченных из генеральной совокупности. Обычно имеется подробная информация об объектах из выборки, а не из генеральной совокупности. Существует много различных способов построения выборки. Каждый способ имеет свои преимущества для определенных целей. Рассмотрим несколько примеров генеральных совокупностей и выборок.

[1] Такое пояснение необходимо, поскольку может не существовать выборки, которая бы *точно* представляла совокупность. Например, это может быть в том случае, если каждый член совокупности уникален.

1. *Генеральная совокупность*: примерно 733 000 жителей г. Талса, штат Оклахома, в котором ваша фирма решила открыть мексиканский ресторан быстрого обслуживания.

   а) Выборку можно построить, наняв людей, которые будут дежурить в местном торговом центре и опрашивать каждого 35-го покупателя. Такая выборка будет содержать информацию о покупателях, но информация об остальной части генеральной совокупности будет отсутствовать.

   б) Другой метод построения выборки — провести опрос по телефону каждого 2000-го жителя города, взяв номера из телефонного справочника. Такая систематическая выборка будет содержать определенную информацию о людях, которые находятся дома и отвечают на телефонные звонки.

   в) Еще один метод построить выборку может заключаться в том, чтобы опросить тех, кто выходит из местного ресторана Мак-Дональдс. Такая выборка даст информацию о группе людей, посещающих рестораны быстрого питания.

2. *Генеральная совокупность*: 826 ящиков с различным компьютерным оборудованием, только что поступивших к вам на склад. Вы хотите проверить на месте содержимое отдельных ящиков, чтобы убедиться, насколько оно соответствует накладной.

   а) Удобный способ заключается в том, чтобы взять 10 ближайших ящиков и проверить их содержимое. Но такая выборка вряд ли будет репрезентативной. К тому же, если ваши поставщики разгадают этот метод отбора, то вы вряд ли сможете извлечь пользу из такой выборки.

   б) Можно подойти к осуществлению выборки иначе: выбрать для проверки три больших, три средних и три небольших по размеру ящика. На первый взгляд, это некоторое расширение метода отбора, но такой вариант вообще может не дать желаемого результата — выборка может оказаться нерепрезентативной (например, почти все ящики могут оказаться больших размеров).

   в) Существует еще один вариант — взять накладную и случайно отобрать ящики для проверки из перечня, указанного в накладной. Затем следует найти и вскрыть отобранные ящики. Это будет наиболее подходящая выборка. Начав с накладной, вы убедитесь в правильности этого документа. Случайность отбора гарантирует, что ваши поставщики не смогут предугадать, какие именно ящики вы будете проверять.

3. *Генеральная совокупность*: ваши поставщики (численностью 598). Вы обдумываете новую систему снабжения, которая предполагает более высокую оплату поставок с одновременным повышением качества и уменьшением времени реагирования на заказ.[2] Эта система будет эффективна только в том случае, если достаточное количество поставщиков будет заинтересовано в ней.

   а) Выборка может состоять из пяти ваших основных поставщиков. Конечно, важно рассмотреть эти крупные фирмы, но стоит также включить в выборку и других поставщиков.

   б) Другую выборку можно получить, полностью доверив право выбора одному из ваших подчиненных (например, направив этому сотруднику сопроводительную записку следующего содержания: "Пожалуйста, составьте список из 10 поставщиков, которые могут работать по системе снабжения "точно во время"). Однако в таком случае вы не будете знать, на основе каких критериев проводился отбор. Можно предположить, что ваш сотрудник будет использовать "наиболее быструю" или "наиболее подходящую" выборку, но это не означает, что полученная выборка будет репрезентативной.

   в) Можно также построить выборку, включив в нее пять ваших ключевых поставщиков и еще 10 поставщиков, отобранных на основе определенных критериев (скажем, отобранных вашим сотрудником на основании следующего задания: "Пожалуйста, составьте список из 10 не основных поставщиков, используя таблицу случайных чисел"). Это будет удачная выборка, поскольку она будет включать как всех наиболее важных поставщиков, так и часть не основных поставщиков.

[2] Эта система похожа на метод организации поставок "точно во время". Вместо создания запасов сырья, хранение которого требует дополнительных затрат и складских помещений, сырье и материалы по такой схеме поступают непосредственно на завод в нужное место и именно тогда, когда они необходимы в производстве.

Что такое репрезентативная выборка

Процесс построения выборки показан на рис. 8.1.1. Из большей по размеру генеральной совокупности извлекается выборка для проведения измерений и подробного анализа (в русском языке словом "выборка" обозначают как сам процесс отбора, так и результат отбора. Смысл слова "выборка", как правило, ясен из контекста. — *Прим. ред.*). При этом предполагается, что выборка является **репрезентативной**. Это означает, что каждое свойство (или комбинация свойств) и в выборке, и в генеральной совокупности имеет одинаковые частоты. О выборке, которая не является репрезентативной, говорят, что она имеет **смещение**. Например, если в выборке доля мужчин больше, чем в генеральной совокупности, то можно сказать, что выборка имеет смещение по полу или что выборка смещена в сторону мужчин.

Поскольку каждый объект может быть уникальным, выборки, которая будет полностью репрезентативной, может не существовать. Как получить достаточно репрезентативную выборку? Если *не* производить целенаправленный отбор, основываясь на некоторой измеряемой характеристике, то случайно взятая статистическая выборка будет свободной (в среднем) от смещений и поэтому репрезентативной (в среднем). Более того, специально введенная в процесс отбора случайность позволит формулировать вероятностные суждения о результатах отбора (например, вести речь о *доверительных интервалах*, как будет показано в следующей главе). Тщательно построенная таким образом статистическая выборка будет почти репрезентативной, *и* вы сможете оценить, насколько она репрезентативна.

Определив для решаемой задачи генеральную совокупность, необходимо выяснить, как с этой совокупностью работать практически. Для практической работы необходимо иметь **основу генеральной совокупности**, которая даст возможность

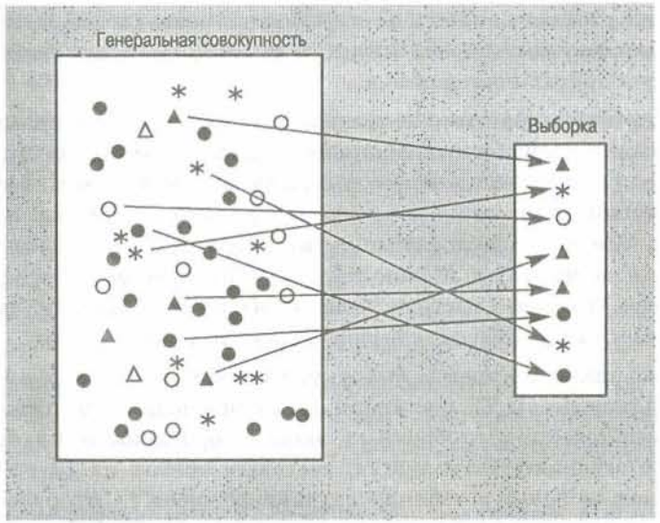


Рис. 8.1.1. Выборка представляет собой набор, извлеченный из большей по размеру генеральной совокупности. Данная выборка будет, по-видимому, достаточно, но не полностью репрезентативной, так как ни один из двух неокрашенных треугольников не попал в выборку

обращаться к отдельным элементам по номерам. В нашем случае основа может иметь вид списка объектов генеральной совокупности, которым присвоены номера от 1 до N, где N — число объектов генеральной совокупности. Например, чтобы получить доступ к объекту генеральной совокупности с номером 137, следует обратиться к списку и в соответствующей строке найти информацию, определяющую объект (например, имя, "серийный номер" или "номер покупателя").

Существуют два основных типа выборки. После того как объект извлечен из генеральной совокупности для включения в выборку, его либо возвращают обратно в генеральную совокупность (и тогда он может попасть в эту же выборку повторно), либо не возвращают. **Выборка без возврата** имеет место, когда любой объект не может попасть в выборку более одного раза, т.е. когда все объекты выборки всегда разные. **Выборка с возвратом** имеет место, если объект генеральной совокупности может попасть в выборку более одного раза. Следует отметить, что эти свойства определяются *процессом*, использованным для осуществления выборки, а не *результатами* этого процесса. Если использовать эти два метода для небольшой выборки, извлеченной из большой генеральной совокупности, то различия будут незначительными. В данной книге мы будем работать главным образом с выборками, содержащими различные объекты, т.е. с выборками *без возврата*.

Мы будем использовать следующие обозначения для количества объектов генеральной совокупности (свойство генеральной совокупности) и для количества объектов, извлекаемых для выборки (это количество зависит от того, как много объектов вы решите отобрать).

Система обозначений для количества объектов

N — размер генеральной совокупности;

n — размер выборки.

Выборка, которая включает полную генеральную совокупность (т.е. такая, что $N = n$), называется **переписью**. Но даже если вы можете изучить всю генеральную совокупность, нужно подумать, стоит ли это делать. Сравнивая затраты и преимущества, можно прийти к выводу, что не имеет смысла тратить время и усилия на изучение всех объектов генеральной совокупности.

Параметры выборки и параметры генеральной совокупности

Параметром выборки (или **выборочным параметром**, или просто **статистикой**) называют показатель (число), вычисленный на основе данных выборки. В качестве примера можно привести выборочное среднее, медиану, стандартное отклонение выборки и процентили. Статистика является *случайной величиной*, так как в ее основе лежат данные, полученные путем случайного отбора, который, в свою очередь, может рассматриваться как случайный эксперимент. Поэтому статистика является *известной* и *случайной* величиной.

Параметр генеральной совокупности (или просто **параметр**) — это показатель (число), вычисленный для всей генеральной совокупности. В качестве примера можно привести среднее и стандартное отклонения генеральной совокупности. Параметр является *фиксированным числом*, так как при его вычислении отсутствует случайность. Однако обычно у нас нет данных обо всей генеральной совокупности. Поэтому параметр является *неизвестной* и *фиксированной* величиной.

Часто существует естественное соответствие между статистиками и параметрами. Для каждого параметра совокупности (показателя, значение которого хотелось бы знать, но которое точно неизвестно) существует выборочная статистика, рассчитанная на основе данных, представляющих наилучшую доступную информацию о неизвестном параметре. Такую выборочную статистику называют **оценочной функцией** параметра генеральной совокупности, а ее фактическое значение, рассчитанное из данных выборки, называют **оценкой** параметра совокупности. Например, среднее выборки является оценочной функцией среднего совокупности, и в конкретном случае оценка может быть равна 18,3. **Ошибкой оценки** называют разность между оценочной функцией (или оценкой) и параметром генеральной совокупности; ошибка оценки обычно неизвестна.

Несмещенная оценка не является систематически слишком завышенной или слишком заниженной в сравнении с соответствующим параметром генеральной совокупности. Такое свойство представляется желательным для оценки. Формально оценка является несмещенной, если ее среднее (среднее ее выборочного распределения) равно соответствующему параметру генеральной совокупности.

Большинство часто используемых оценок являются несмещенными или почти (асимптотически) несмещенными. Например, среднее выборки $\overline{X}$ является несмещенной оценкой среднего генеральной совокупности μ. Конечно, для любого *конкретного* набора данных $\overline{X}$ обычно будет больше или меньше среднего генеральной совокупности, μ. Если многократно повторять процесс извлечения вы-

борки и для каждой выборки вычислять $\overline{X}$, то полученные результаты будут в среднем близки к μ и, следовательно, не будут *систематически* слишком высокими или слишком низкими.

Стандартное отклонение выборки S является (как ни странно) смещенной оценкой стандартного отклонения генеральной совокупности σ, но в то же время асимптотически (приблизительно) несмещенным. Дисперсия выборки S^2 представляет собой несмещенную оценку дисперсии генеральной совокупности σ^2. В случае биномиального распределения выборочная доля p является несмещенной оценкой доли в генеральной совокупности π.

8.2. Случайная выборка

Случайная выборка, или **простая случайная выборка**, строится таким образом, что (1) каждый объект генеральной совокупности имеет *одинаковую вероятность быть отобранным* и (2) объекты *отбирают независимо* друг от друга. Если элементы совокупности имеют равную вероятность быть извлеченными, то полученная случайная выборка будет достаточно хорошей и, насколько это возможно, несмещенной. Независимость отбора обеспечивает сбор максимально возможного объема независимой информации.[3] Поскольку индивидуальные вкусы и человеческий фактор исключены из процесса отбора, у полученной таким образом выборки будет больше шансов быть репрезентативной, чем у той "произвольной" выборки, которую вы можете поручить сделать кому-либо.

Другой возможный и эквивалентный способ определить *случайную выборку* заключается в том, чтобы сказать, что это выборка, выбранная случайно из множества всех возможных выборок такого объема, которые можно было бы извлечь из генеральной совокупности. С таким определением работать сложнее, так как число возможных выборок может быть огромно. Например, существует 17 310 309 456 440 различных выборок объемом $n = 10$ объектов, которые можно извлечь из генеральной совокупности, содержащей 100 объектов.[4] Однако такое определение ясно демонстрирует, что в процессе построения случайной выборки мы не отдаем предпочтение ни одной из потенциально возможных выборок.

Насколько случайная выборка лучше произвольной? При извлечении случайной выборки у вас есть гарантия, что математическая статистика на вашей стороне. Вы не просто "надеетесь на лучшее", а получаете настоящую гарантию, что выборка является репрезентативной, по крайней мере в среднем, для всех характеристик генеральной совокупности (даже для тех характеристик, которые еще не встретились, и тех, которые трудно или невозможно измерить!). Кроме того, случайная выборка закладывает основу для корректности заключений (статистических выводов) относительно генеральной совокупности, которые могут быть сделаны исходя из данных этой выборки. С другой стороны, например, если вы извлекаете неслучайную выборку, которая должна быть репрезентативной в отношении (a) количества мужчин и женщин, (b) семейного положения и (c) дохода, то результирующая выборка может быть совершенно отличной от генеральной совокупности по таким важным характеристикам, как использование Internet или желание делать заказы по каталогам. Это легко может привести к неудачным бизнес-решениям, так как вы не использовали случайную выборку.

[3] Хотя независимость является формальным понятием, она имеет важные практические последствия. Ниже приведен пример, который поможет понять проблемы, возникающие, когда элементы совокупности отбираются *не* независимо: в больнице с 20-ю палатами, в каждой из которых находится по 50 пациентов, следует взять выборку (но *не* случайную выборку) путем случайного выбора палаты и проведения в ней опроса всех пациентов. Следует отметить, что каждый пациент имеет равный шанс быть опрошенным (один из двадцати). Однако, так как вместо независимого отбора пациентов респонденты опроса отбираются целой группой, в этой выборке будет отсутствовать важная информация обо всех пациентах больницы.

[4] Количество различных выборок без возврата определяется формулой $\binom{N}{n} = N!/[n!(N-n)!]$, в которой вы можете узнать часть формулы для биномиального распределения вероятности.

Извлечение случайной выборки

Одним из способов извлечения случайной выборки является применение таблицы случайных чисел для получения номера каждого отобранного объекта генеральной совокупности. Сам объект затем находят в основе выборки (это является главным назначением основы выборки: переход от номера непосредственно к самому элементу генеральной совокупности). **Таблица случайных чисел** представляет собой организованную в виде таблицы последовательность цифр, в которой каждая из цифр от 0 до 9 встречается независимо друг от друга с вероятностью 1/10. Ниже приведена подробная схема извлечения без возврата случайной выборки размером n.

Извлечение случайной выборки без возврата

1. Составьте основу выборки таким образом, чтобы все элементы генеральной совокупности были пронумерованы числами от 1 до N.
2. Выберите точку начала считывания случайных чисел из таблицы. Это необходимо сделать случайным образом, например подбросив монету.
3. Начав с выбранной точки, последовательно считывайте цифры обычным способом (например, слева направо, с переходом на следующую строку).
4. Объедините эти цифры в группы, размер которых равен количеству цифр в числе N. Например, при размере генеральной совокупности N= 5387 считывайте по четыре случайные цифры за раз (так как запись числа 5387 включает четыре цифры). Или, если размер генеральной совокупности N= 3163298 элементов, то объединяйте прочитанные случайные цифры в группы по семь.
5. Выполняйте следующие действия до тех пор, пока не получите выборку из n элементов.
   а) Если вы получили случайное число между 1 и N и элемент с таким номером еще не извлекался, включите его в выборку.
   б) Если полученное случайное число равно 0 или больше N, то отбросьте его, так как для него в основе выборки нет соответствующего элемента генеральной совокупности.
   в) Если получено такое случайное число, что элемент с соответствующим номером уже был извлечен ранее, то отбросьте это число, так как вы осуществляете выборку без возврата.

Например, давайте построим случайную выборку, содержащую 9 потребителей, из списка, в который входят 38 человек. Начнем с числа 69506 таблицы случайных чисел (ряд 11, столбец 3, табл. 8.2.1). Поскольку число $N = 38$ состоит из двух цифр, объединим последовательность случайных чисел в группы, состоящие из двух цифр, следующим образом: 69 50 61 96 10 01 47 99 23 38... . Отбрасываем первые несколько чисел из этой последовательности, так как 69, 50, 61 и 96 больше, чем 38 ($N = 38$). Первым попавшим в выборку будет число 10. Остальные

действия показаны на рис. 8.2.1. Когда число 10 попадет во второй раз, не следует включать его в выборку (так как мы строим выборку без возвратов). Процесс продолжаем до тех пор, пока не будет отобрано $n = 9$ элементов.[5]

Начинаем с использования таблицы случайных чисел:
69506 19610 01479 92338 55140 81097 73071 61544 85356 51400

Объединяем цифры в группы по две (поскольку число 38 состоит из двух цифр):
69 50 61 96 10 01 47 99 23 38 55 14 08 10 97 73 07 16 15 44 85 35 65 14 00

Исключаем числа, которые больше 38 или меньше 1:
10 01 23 38 14 08 10 07 16 15 35 14

Исключаем числа, которые уже встречались ранее:
10 01 23 38 14 08 07 16 15 35

Выбираем первые девять чисел:
10 1 23 38 14 8 7 16 15

Рис. 8.2.1. Извлечение без возврата случайной выборки размером $n = 9$ элементов из генеральной совокупности размером $N = 38$ элементов. Случайные числа объединяются в группы из двух цифр (так как число 38 имеет две цифры) начиная с ряда 11, столбца 3 табл. 8.2.1. Числа больше 38 или меньше 1 исключаются. Число 10, когда оно встречается во второй раз, также исключается. Процедура завершается, когда получено n единиц

Таблица 8.2.1. Таблица случайных чисел

	1	2	3	4	5	6	7	8	9	10
1	51449	39284	85527	67168	91284	19954	91166	70918	85957	19492
2	16144	56830	67507	97275	25982	69294	32841	20861	83114	12531
3	48145	48280	99481	13050	81818	25282	66466	24461	97021	21072
4	83780	48351	85422	42978	26088	17869	94245	26622	48318	73850
5	95329	38482	93510	39170	63683	40587	80451	43058	81923	97072
6	11179	69004	34273	36062	26234	58601	47159	82248	95968	99722
7	94631	52413	31524	02316	27611	15888	13525	43809	40014	30667
8	64275	10294	35027	25604	65695	36014	17988	02734	31732	29911
9	72125	19232	10782	30615	42005	90419	32447	53688	36125	28456
10	16463	42028	27927	48403	88963	79615	41218	43290	53618	68082
11	10036	66273	69506	19610	01479	92338	55140	81097	73071	61544
12	85356	51400	88502	68267	73943	25828	38219	13268	09016	77465
13	84076	82087	55053	75370	71030	92275	55497	97123	40919	57479
14	76731	39755	78537	51937	11680	78820	50082	56068	36908	55399

[5] Если вы берете выборку *с* возвратом, последовательность ваших действий будет такой же, за исключением п. 5,а (где вы должны включить в выборку любые числа из интервала от 1 до N) и п. 5,в (который следует проигнорировать).

Окончание табл. 8.2.1

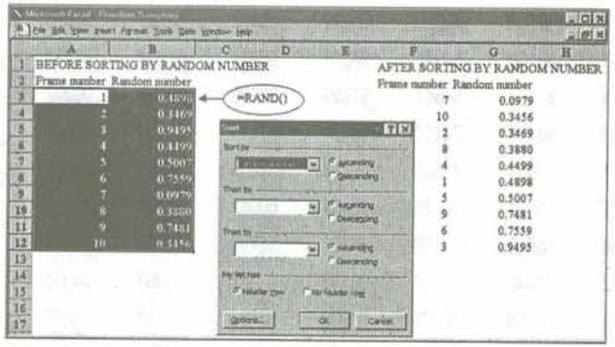
	1	2	3	4	5	6	7	8	9	10
15	19032	73472	79399	05549	14772	32746	38841	45524	13535	03113
16	72791	59040	61529	74437	74482	76619	05232	28616	98690	24011
17	11553	00135	28306	65571	34465	47423	39198	54456	95283	54637
18	71405	70352	46763	64002	62461	41982	15933	46942	36941	93412
19	17594	10116	55483	96219	85493	96955	89180	59690	82170	77643
20	09584	23476	09243	65568	89128	36747	63692	09986	47687	46448
21	81677	62634	52794	01466	85938	14565	79993	44956	82254	65223
22	45849	01177	13773	43523	69825	03222	58458	77463	58521	07273
23	97252	92257	90419	01241	52516	66293	14536	23870	78402	41759
24	26232	77422	76289	57587	42831	87047	20092	92676	12017	43554
25	87799	33602	01931	66913	63008	03745	93939	07178	70003	18158
26	46120	62298	69129	07862	76731	58527	39342	42749	57050	91725
27	53292	55652	11834	47581	25682	64085	26587	92289	41853	38354
28	81606	56009	06021	98392	40450	87721	50917	16978	39472	23505
29	67819	47314	96988	89931	49395	37071	72658	53947	11996	64631
30	50458	20350	87362	83996	86422	58694	71813	97695	28804	58523
31	59772	27000	97805	25042	09916	77569	71347	62667	09330	02152
32	94752	91056	08939	93410	59204	04644	44336	55570	21106	76588
33	01885	82054	45944	55398	55487	56455	56940	68787	36591	29914
34	85190	91941	86714	76593	77199	39724	99548	13827	84961	76740
35	97747	67607	14549	08215	95408	46381	12449	03672	40325	77312
36	43318	84469	26047	86003	34786	38931	34846	28711	42833	93019
37	47874	71365	76603	57440	49514	17335	71969	58055	99136	73589
38	24259	48079	71198	95859	94212	55402	93392	31965	94622	11673
39	31947	64805	34133	03245	24546	48934	41730	47831	26531	02203
40	37911	93224	87153	54541	57529	38299	65659	00202	07054	40168
41	82714	15799	93126	74180	94171	97117	31431	00323	62793	11995
42	82927	37844	74411	45887	36713	52339	68421	35968	67714	05883
43	65934	21782	35804	36676	35404	69987	52268	19894	81977	87764
44	56953	04356	68903	21369	35901	86797	83901	68681	02397	55359
45	16278	17165	67843	49349	90163	97337	35003	34915	91485	33814
46	96339	95028	48468	12279	81039	56531	10759	19579	00015	22829
47	84110	49661	13988	75909	35580	18426	29038	79111	56049	96451
48	49017	60748	03412	09880	94091	90052	43596	21424	16584	67970
49	43560	05552	54344	69418	01327	07771	25364	77373	34841	75927
50	25206	15177	63049	12464	16149	18759	96184	15968	89446	07168

Извлечение выборки методом перемешивания генеральной совокупности

Другой способ извлечения случайной выборки из генеральной совокупности можно легко осуществить с помощью компьютерной программы работы с электронными таблицами. Идея заключается в том, чтобы перемешать элементы генеральной совокупности случайным образом и затем отобрать в выборку необходимое количество элементов. Это похоже на то, как тасуют колоду игральных карт, чтобы затем сдать необходимое для игры количество карт.

В одном столбце располагают числа от 1 до N; обычно есть соответствующая команда, которая позволяет создать такой столбец автоматически. Следующий столбец с помощью генератора случайных чисел заполняют равномерно распределенными случайными числами из интервала от 0 до 1 таким образом, чтобы эти случайные числа находились рядом с числами первого столбца. На следующем шаге оба столбца сортируют таким образом, чтобы упорядочить числа во втором столбце. В результате все элементы генеральной совокупности будут перемешаны (перетасованы) случайным образом. Наконец, чтобы осуществить выборку, берут первые n элементов из этой перемешанной генеральной совокупности.

Чтобы построить с помощью Excel случайную выборку объемом $n = 3$ из генеральной совокупности размером $N = 10$, в верхнюю ячейку столбца случайных чисел следует ввести формулу = RAND() (=СЛЧИС()), нажать клавишу <ENTER> и затем скопировать результат вниз по столбцу, чтобы получить столбец случайных чисел. Выделив оба столбца (с номерами элементов в основе выборки и со случайными числами), выполните команду Data⇨Sort (Данные⇨Сортировка) из меню Excel, чтобы выполнить сортировку строк, используя значения из столбца со случайными числами. После этого числа в первом столбце будут упорядочены случайным образом, и для получения искомой случайной выборки достаточно будет взять первых три номера элементов из основы выборки (т.е. три первых числа из первого столбца). В данном примере в выборку попали элементы с номерами 7, 10 и 2.



Полученная таким образом случайная выборка будет обладать теми же свойствами, что и выборка, построенная с использованием таблицы случайных чисел.

Пример. Аудит

В годовом отчете за 1998 финансовый год корпорация *Microsoft* указала, что ею получен доход в размере 14,5 млрд. дол., а чистая прибыль составила 4,5 млрд. дол. Число отдельных сделок огромно, и, чтобы быть уверенным в правильности этих цифр, отчетность тщательно изучается. Ниже приведено мнение аудиторской фирмы *Deloitte & Touche LLP*[6].

"По нашему мнению, финансовые отчеты представлены надлежащим образом. Во всех отношениях финансовое состояние корпорации *Microsoft* и ее филиалов за период с 30 июня 1997 г. по 30 июня 1998 г., а также результаты ее операций и движение наличных денег за каждый из трех финансовых годов по состоянию на 30 июня 1998 г. соответствуют общепринятым стандартам финансового учета".

Далее аудиторская фирма приводит следующие обоснования (ниже представлена часть) своего мнения.

"Нами проведена аудиторская проверка в соответствии с общепринятыми стандартами аудита. Эти стандарты требуют, чтобы мы планировали и проводили аудит таким способом, который дает достаточную гарантию, что в финансовых отчетах отсутствуют ошибки и неточности. Аудит заключается в изучении и проверке доказательств, подтверждающих (или раскрывающих) заявленные в финансовых отчетах суммы... Мы уверены, что наш аудит обеспечивает разумную основу для сделанного заключения".

В такого рода аудиторских проверках присутствует статистика, так как необходимо анализировать данные о большом количестве сделок. Тщательно проверяются все крупные сделки. В то же время при проверке большого количества небольших сделок многие аудиторы полагаются на статистическую выборку.[7]

Рассмотрим конкретный список сделок (вероятно, один из многих таких списков), пронумерованных числами от 1 до 7329. Вам необходимо построить случайную выборку из 20 финансовых счетов, приняв за начальную точку отсчета ряд 23, столбец 8 в таблице случайных чисел. Расположив случайные числа в группы из четырех цифр и заключив в квадратные скобки большие числа, которые необходимо исключить из рассмотрения, получим начальный список:

2387 0784 0241 [7592] 6232 [7742] 2762 [8957] 5874 2831 [8704] 7200 [9292] 6761 2017
4355 4877 [9933] 6020 1931 6691 3630 0803 [7459] 3939 0717 [8700] 0318 1584 6120 ...

Выбрав первые $n = 20$ чисел из этого списка, получим выборку, включающую следующие номера сделок:

2387 784 241 6232 2762 5874 2831 7200 6761 2017 4355 4877 6020
1931 6691 3630 803 3939 717 318

Расположив их по возрастающей, составим окончательный список сделок, образующих необходимую нам выборку:

241 318 717 784 803 1931 2017 2387 2762 2831 3630 3939 4355 4877
5874 6020 6232 6691 6761 7200

Затем следует подробно рассмотреть эти сделки и проверить правильность соответствующих финансовых документов. Информацию, полученную с помощью выборки из этого списка сделок, следует объединить с другой информацией, полученной путем выборки из других списков сделок, а также с информацией, полученной путем тщательного изучения крупных и важных сделок.

Пример. *Пробное исследование фирм, занимающихся выпуском бумаги и лесоматериалов*

У вас есть новое потенциально очень полезное изделие для компаний, занимающихся выпуском бумаги и лесоматериалов. Чтобы заранее выработать маркетинговую стратегию, вы решили собрать информацию об этих фирмах. Проблема заключается еще и в том, что изделие до конца не разработано, и вы даже не

[6] Взято из "*Microsoft 1998 Annual Report*" по адресу: http://www.microsoft.com/msft/ar98.htm.

[7] Обзор методов, которые могут помочь при проведении аудита, приведен в работе *A. J. Wilburn, Practical Sampling for Auditors. New York: Marcel Dekker, 1984.*

знаете точно, как собрать информацию! Поэтому вы решаете провести **пробное (пилотное) исследование**, которое является мини-версией настоящего исследования и предназначено для того, чтобы определить возможные проблемы и решить их до проведения настоящего полномасштабного исследования. Для вашего пробного исследования вы решили взять только три из этих фирм ($n = 3$), отобранных случайным образом.

Вначале вы строите основу выборки, как это показано в табл. 8.2.2. Взяв в качестве начальной точки ряд 39, столбец 6, считываем парами цифры из таблицы случайных чисел (поскольку число $N = 32$ состоит из двух цифр). Заключив в квадратные скобки числа, подлежащие исключению, получаем следующий начальный список:

[48], [93], [44], 17, 30 [47], [83], 12, [65], 31, 02, 20 [33], [79], 11, [93], 22, [48], [71], [53], ...

Взяв первых $n = 3$ числа из этого списка, получаем, что в выборку войдут фирмы со следующими номерами:

17, 30, 12

Затем, обратившись к основе выборки, по номерам определяем названия фирм. Расположив названия в алфавитном порядке, получаем для пробного исследования такую выборку из $n = 3$ следующих фирм: Georgia-Pacific, Longview Fibre и Wesvaco.

Таблица 8.2.2. Основа выборки

1	Avery Dennison	17	Longview Fibre
2	Bemis	18	Louisiana-Pacific
3	Boise Cascade	19	Manville
4	Bowater	20	Mead
5	Champion International	21	Ply Gem Industries
6	Chesapeake	22	Potlatch
7	Consolidated Papers	23	Rayonier
8	Federal Paper Board	24	Scott Paper
9	First Brands	25	Sonoco Products
10	Fort Howard	26	Stone Container
11	Gaylord Container	27	Temple-Inland
12	George-Pacific	28	Union Camp
13	International Paper	29	Universal Forest Products
14	James River	30	Westvaco
15	Jefferson Smurfit	31	Weyerhaeuser
16	Kimberly-Clark	32	Willamette Industries

8.3. Выборочное распределение и центральная предельная теорема

Любая статистика, вычисленная на основе случайной выборки, будет иметь распределение вероятностей, которое называют **выборочным распределением** этой статистики. Знание выборочного распределения дает возможность перейти

от информации о выборке (которая у вас уже есть) к информации о генеральной совокупности (которую вам хотелось бы иметь). К счастью, во многих случаях выборочное распределение статистики, такой как, например, среднее выборки, близко к нормальному даже тогда, когда распределение для отдельных объектов отличается от нормального. Этот результат, который называют *центральной предельной теоремой*, упрощает статистический вывод, поскольку известно, как вычислять вероятности для нормального распределения.

Когда вы заканчиваете отчет о результатах опроса случайно отобранных покупателей и видите, что они готовы тратить на покупку бакалейных товаров в среднем $21,26 за один визит в магазин, число 21,26 для вас не выглядит случайным. *Но результат вашего обследования является случайным.* Будьте осторожны. Число 21,26 само по себе не случайно. Однако оно отражает "средние расходы на бакалейные товары за один визит в магазин для случайно выбранных покупателей", а это есть случайная переменная. Если взглянуть на число 21,26 с этой точки зрения, становится ясным, почему оно является случайным: в результате выполнения случайного эксперимента каждый раз опрашивается новая случайная выборка покупателей и, следовательно, каждый раз будет разный результат.

То, как мы *думаем* о статистике, может коренным образом отличаться от того, как мы *работаем* с ней. Чтобы разобраться в приведенных понятиях, представьте, что исследование повторяется много раз. Это необходимо, чтобы понять, откуда возникает случайность; в конце концов, если исследование проведено только один раз, результаты будут просто конкретными числами. Однако необходимо также понимать, что вследствие ограничений реальной жизни вы действительно проводите исследование, как правило, только один раз. Предположение о многократном повторении исследования представляет собой лишь способ понять имеющийся фактический результат, поместив его в контекст всех других возможных результатов. Учитывая это, внимательно изучите идеи выборочного распределения, представленные на рис. 8.3.1.

Существуют две причины, по которым нормальное распределение столь важно. Во-первых, многие наборы данных подчиняются нормальному распределению (хотя в бизнесе для исключения асимметрии значения часто приходится логарифмировать). Во-вторых, даже если распределение набора данных не является нормальным, распределение *среднего* или *суммы* чисел из такого набора будет близким к нормальному.

Важно отличать *отдельное* значение от среднего или суммы, которые объединяют много значений. Несмотря на то что отдельные значения подчиняются какому-то неизвестному нам закону распределения, процедура объединения многих значений для вычисления среднего или суммы приводит к нормальному распределению.

Чтобы понять это, следует осознать, что процесс получения случайной выборки и вычисление для этой выборки среднего является некоторым случайным экспериментом, а среднее такой выборки представляет собой случайную переменную. Поэтому имеет смысл говорить, является распределение нормальным или нет, только в отношении распределения среднего или суммы. Теперь, имея дело с вероятностями, а не со статистиками, можно легко представить многократно повторяющийся процесс такого сбора данных, который приводит к мно-

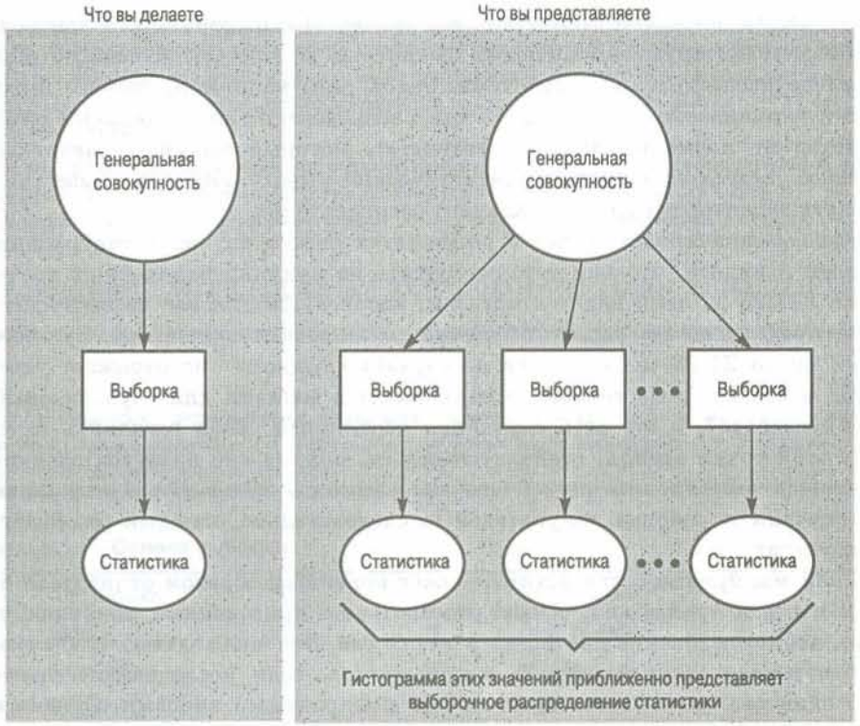


Рис. 8.3.1. В предположении, что исследование повторяется многократно, выборочное распределение статистики соответствует гистограмме значений статистики. В действительности, конечно, имеется только одна выборка и одно значение статистики. Это единственное значение интерпретируют с учетом всех остальных результатов, которые могли бы иметь место, как это представлено выборочным распределением

гократным наблюдениям среднего. Гистограмма этих наблюдений представляет (приближенно) распределение среднего.

Центральная предельная теорема утверждает, что для случайной выборки объемом n элементов из генеральной совокупности справедливы следующие утверждения.

1. С увеличением n распределение как *среднего*, так и *суммы* все больше приближается к нормальному.
2. Средние и стандартные отклонения распределений среднего и суммы вычисляют по приведенным ниже формулам, где μ — среднее и σ — стандартное отклонение элементов генеральной совокупности.

Центральная предельная теорема дает всю информацию, необходимую для вычисления вероятностей для суммы и среднего случайной выборки. Если значение n достаточно велико, то можно принять, что имеет место нормальное распределение, и использовать стандартные таблицы для нормального распределения вероят-

Среднее и стандартное отклонение для средних и сумм

	Случайная переменная	
	среднее	общая сумма
Среднее	$\mu_{\bar{X}} = \mu$	$\mu_{sum} = n\mu$
Стандартное отклонение	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$	$\sigma_{sum} = \sigma\sqrt{n}$

ностей.[8] Для стандартизации значений можно использовать соответствующие среднее и стандартное отклонение из приведенной выше таблицы. Поэтому все, что надо уметь, — так это вычислять вероятности для нормального распределения!

На рис. 8.3.2 приведена гистограмма объемов продаж ста наиболее известных промышленных корпораций США.[9] Из нее видно, что это распределение достаточно асимметрично. На рис. 8.3.3 показано распределение средних значений для пяти фирм, взятых из этого списка 100 раз (т.е. средние значения объемов продаж были вычислены 100 раз для пяти случайно отобранных из этого списка фирм), представляющее выборочное распределение средних для пяти фирм. Распределение все еще асимметрично, но уже в меньшей мере, и ближе к нормальному, чем распределение объемов продаж отдельных фирм. На рис. 8.3.4 использован больший размер выборки, $n = 25$. Следует отметить, что значения среднего остаются теми же, тогда как стандартные отклонения становятся все меньше в соответствии с правилом "деления на корень квадратный из n". Также обратите внимание на переход в диаграммах от асимметричного к нормальному распределению.

Как работает центральная предельная теорема? Идея заключается в том, что экстремальные значения данных взаимно усредняются. Формирующая асимметрию затянутая вправо часть распределения на рис. 8.3.2 перемещается внутрь, так как очень крупные фирмы усредняются с другими фирмами. Этим и объясняется уменьшение асимметрии.

Почему распределение стремится к нормальному? Исчерпывающий ответ на этот вопрос связан с боле глубокими результатами математической статистики, которые в этой главе не представлены. Однако это общий теоретический результат, для которого важно, чтобы значение σ было конечным и не равным нулю.

---

[8] Что означает выражение "достаточно велико"? Если распределение элементов не слишком асимметрично, значение $n = 30$ является обычно достаточным. Однако, если распределение сильно асимметрично или имеет большие отклонения, n, вероятно, должно быть значительно больше. Если распределение приближается к нормальному, тогда n может быть значительно меньше 30, скажем, 20, 10 или даже 5. Конечно, если первоначально распределение было нормальным, тогда достаточно, чтобы n равнялось 1.

[9] Данные взяты из "*The Fortune 500 Largest U.S. Industrial Corporations*", *Fortune*, April 20, 1992, pp. 220-222.

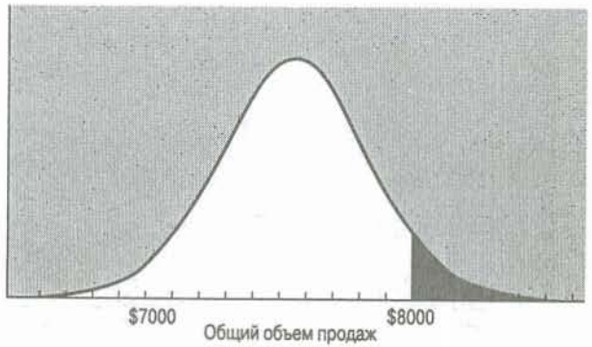


Рис. 8.3.5. Вероятность того, что общий объем продаж превысит $8000 составляет 0,0436. Среднее и стандартное отклонения были вычислены на основе центральной предельной теоремы

Пример. *Стабильность параметров при производстве жевательной резинки*

Это нормально, если каждый отдельный пакетик жевательной резинки не весит точно 0,20 унции, как это написано на упаковке, при условии, что средний вес пакетика не слишком мал (чтобы избежать потери репутации фирмы, не говоря уже об исках потребителей и правительственных организаций) и не слишком велик (чтобы избежать лишних затрат). Из опыта вы знаете, что на вашем производстве вес отдельных кусочков жевательной резинки имеет стандартное отклонение в 0,074 унции, что характеризует изменчивость относительно среднего значения веса, равного 0,201 унции. Любую упаковку из 30 штук жевательных резинок, в которой средний вес одной жевательной резинки ниже 0,18 унции, выбраковывают. Какая часть упаковок будет выбракована таким образом?

Допустим, что кусочки жевательной резинки изготавливают независимо (что не совсем разумно, так как производственная проблема может затрагивать одновременно много кусочков).

Вначале следует определить значения среднего и стандартного отклонения среднего для $n = 30$ кусочков, где каждый кусочек имеет средний вес $\mu = 0{,}201$ унции со стандартным отклонением $\sigma = 0{,}074$.

$$\mu_{(\text{средний вес})} = \mu = 0{,}201 \text{унции}$$

$$\sigma_{(\text{средний вес})} = \frac{\sigma}{\sqrt{30}} = \frac{0{,}074}{5{,}477226} = 0{,}01351 \text{унции}.$$

Затем преобразуем 0,18 унций в нормированное значение:

$$z = \text{предел для нормированного среднего веса} = \frac{0{,}18 - \mu_{(\text{средний вес})}}{\sigma_{(\text{средний вес})}} = \frac{0{,}18 - 0{,}201}{0{,}01351} = -1{,}55.$$

И, наконец, получаем, что вероятность выбраковки упаковки — это вероятность того, что нормированная и нормально распределенная переменная меньше, чем –1,55. По стандартной таблице нормального распределения вероятностей находим, что вероятность равна 0,06, т.е. около 6% упаковок будет выбраковано. Эта вероятность показана на рис. 8.3.6.

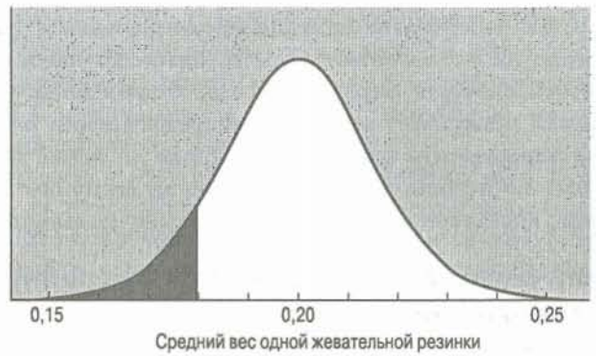


Рис. 8.3.6. Вероятность того, что средний вес одной жевательной резинки будет ниже 0,18 унции, составляет 0,06. Среднее и стандартное отклонения вычислены на основе центральной предельной теоремы

8.4. Стандартная ошибка как *оценка* стандартного отклонения

К сожалению, в реальной жизни обычно нет возможности работать непосредственно с выборочным распределением, поскольку его параметры определяются свойствами всей генеральной совокупности, а информация имеется только о выборке. Каждое распределение характеризуется стандартным отклонением, поэтому выборочное распределение любой статистики также имеет стандартное отклонение. Если это стандартное отклонение известно, то будет приблизительно известно и то, насколько выборочная статистика отличается от своего среднего значения (соответствующего параметра генеральной совокупности). Это в дальнейшем поможет получить больше информации о совокупности, так как в дополнение к имеющемуся "наилучшему предположению" (вашей статистике) вы узнаете, насколько это предположение является удачным. К сожалению, точное значение стандартного отклонения неизвестно, поскольку оно зависит от генеральной совокупности.

Решение проблемы заключается в использовании информации о выборке, чтобы сделать предположение или оценить стандартное отклонение выборочного распределения статистики. Полученное таким образом приближение стандартного отклонения статистики, основанное только на данных выборки, называют **стандартной ошибкой статистики**. Стандартную ошибку интерпретируют так же, как и любое стандартное отклонение. Стандартная ошибка показывает приблизительно, насколько наблюдаемое значение статистики отличается от ее среднего значения. Более точно, стандартная ошибка показывает (приближенно), каким будет стандартное отклонение в том случае, если взять большое количество выборок, определить для каждой из этих выборок среднее и рассмотреть эти выборочные средние как набор данных.

Почему используют два термина (стандартное отклонение и стандартная ошибка), хотя стандартная ошибка является просто одним из видов стандартного отклонения? Это обусловлено, в первую очередь, стремлением подчеркнуть, что *стандартная ошибка* показывает величину неопределенности *итогового значения* (статистики), характеризующего всю выборку. И в то же время термин *стандартное отклонение* обычно используют для обозначения величины изменчивости отдельных *элементов*, характеристики отличия отдельных элементов от среднего.

Насколько отличается среднее выборки от среднего генеральной совокупности

Среднее выборки $\overline{X}$ является статистической величиной, так как его вычисляют на основе данных выборки. **Стандартная ошибка среднего** (или, кратко, просто **стандартная ошибка**) оценивает выборочную изменчивость выборочного среднего, приближенно показывая, насколько выборочное среднее отличается от среднего генеральной совокупности. Из центральной предельной теоремы (раздел 8.3) известно, что стандартное отклонение среднего выборки определяется по формуле $\sigma_{\overline{X}} = \sigma / \sqrt{n}$. В разделе 8.3 сделано допущение, что параметр генеральной совокупности σ известен, поскольку мы работали с вероятностями. Здесь же, когда мы снова работаем со статистиками, значение σ неизвестно, а значит, неизвестно и значение стандартного отклонения среднего выборки. Но есть оценка стандартного отклонения: S, стандартное отклонение выборки, описанное в главе 5. Если заменить σ на S, то в результате получим показатель неопределенности $\overline{X}$. Ниже приведены формулы для определения стандартного отклонения $\overline{X}$ (точное значение) и стандартной ошибки $\overline{X}$ (приблизительное значение, поскольку это значение оценки).

Стандартное отклонение среднего

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Стандартная ошибка среднего

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}.$$

Стандартная ошибка указывает, насколько среднее выборки $\overline{X}$ отличается от среднего генеральной совокупности μ. Поскольку часто $\overline{X}$ — это ваша наилучшая информация о μ, стандартная ошибка грубо показывает, насколько вы ошибаетесь, используя лучшую доступную выборочную информацию (например, среднюю стоимость покупок 100 случайных покупателей) вместо недоступной информации о генеральной совокупности (средняя стоимость покупок всех покупателей города). Размерность у стандартной ошибки среднего та же, что и у исходных данных (доллары, количество миль на один галлон, количество людей и т.п.).

Следует различать S и $S_{\overline{X}}$. Стандартное отклонение S приближенно показывает, насколько отдельные значения отличаются от среднего значения набора данных, в то время как стандартная ошибка $S_{\overline{X}}$ приближенно показывает, насколько среднее $\overline{X}$ отличается от среднего генеральной совокупности μ. Это проиллюстрировано на рис. 8.4.1 и 8.4.2.

Можно ожидать, что в 95% случаев выборочное среднее $\overline{X}$ будет находиться в пределах двух стандартных ошибок от среднего генеральной совокупности. Ниже приведена итоговая таблица, показывающая различие между вариацией отдельных значений и вариацией средних как для генеральной совокупности, так и для выборки.

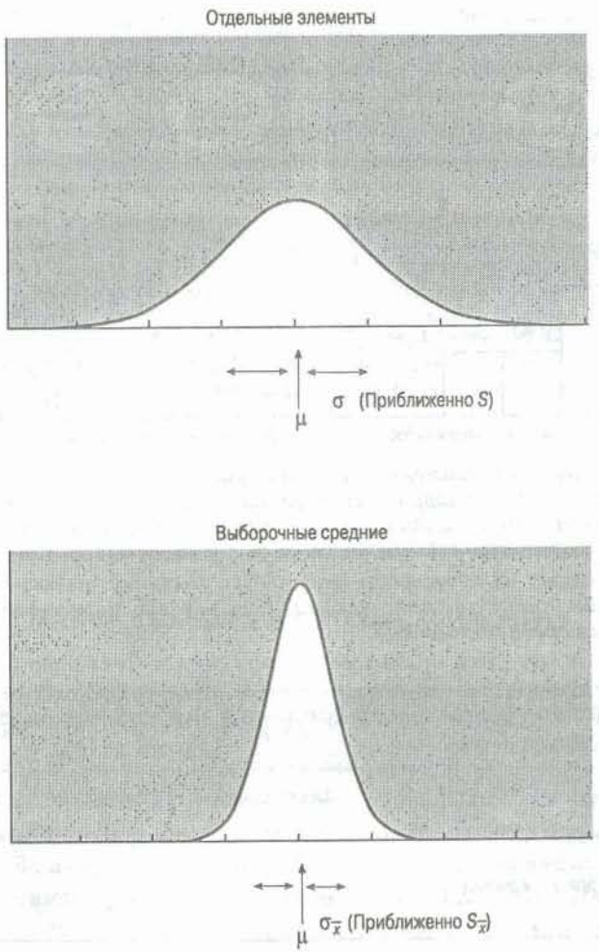


Рис. 8.4.1. У среднего выборки $\overline{X}$ изменчивость меньше, чем у значений элементов X. Стандартная ошибка $S_{\overline{X}}$ меньше, чем стандартное отклонение S, и уменьшается (указывая на большую точность $\overline{X}$) с ростом n; на данном рисунке $n = 4$

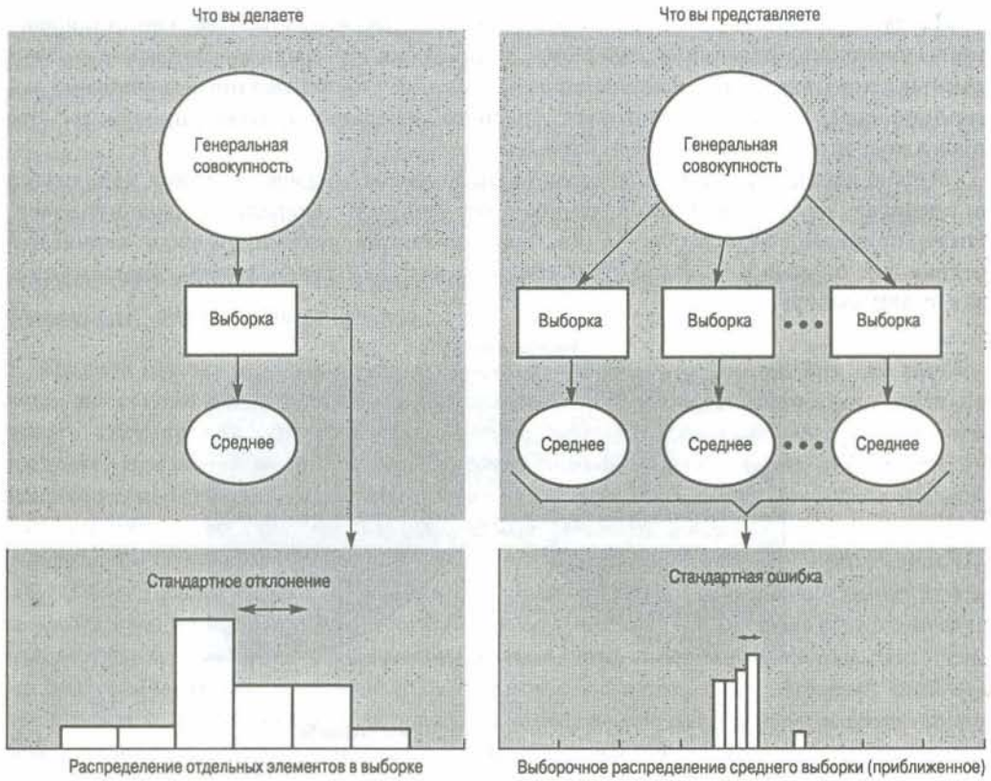


Рис. 8.4.2. Случайные эксперименты и гистограммы результирующих данных для элементов (слева) и выборочных средних (справа). Обратите внимание, насколько стандартная ошибка меньше, чем стандартное отклонение (соотношение приблизительно 1/3 при размере выборки $n = 10$). Стандартное отклонение показывает изменчивость отдельных значений относительно их среднего, тогда как стандартная ошибка показывает изменчивость выборочного распределения среднего выборки

Вариация: отдельные значения и средние, генеральная совокупность и выборка

	Для генеральной совокупности	Для выборки
Вариация отдельных значений	σ	S
Вариация $\overline{X}$, среднего для n значений	$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$	$S_{\overline{X}} = \frac{S}{\sqrt{n}}$

Почему необходимо иметь результат более чем одного наблюдения? А потому, что элементы более изменчивы и случайны, а также менее точны, чем среднее значение выборки. Вот почему стандартную ошибку вычисляют путем деления S на $\sqrt{n}$, и, таким образом, она всегда меньше S при n равном или больше 2.

Почему чем меньше количество элементов, тем больше следует делать выборок? Потому что ошибка $(\overline{X} - \mu)$ обычно уменьшается, когда выборка включает информацию о большем количестве элементов. Стандартная ошибка показывает приближенную величину этой ошибки. Ввиду того что стандартная ошибка тем меньше, чем больше n (при прочих равных условиях), информация о неизвестном μ улучшается с ростом размера выборки, так как значение $\overline{X}$ приближается к значению μ.

Пример. *Поездки за покупками*

Пусть n = 200 случайно выбранных для опроса покупателей утверждают, что в этот день они планируют потратить в среднем $\overline{X}$ = \$19,42 при стандартном отклонении равном S = \$8,63. Другими словами, обычно покупатели планируют потратить на покупки \$19,42, а отдельный покупатель планирует потратить примерно на 8,63 больше или меньше этой суммы. Пока мы имеем не более чем описание опрошенных людей.

Фактически же можно не только дать описание выборочных данных, но и сделать некоторые утверждения о неизвестном среднем генеральной совокупности μ, являющемся средней суммой денег, которые планируют потратить сегодня все покупатели, включая и тех, кого вы не опросили. Стандартная ошибка будет равна

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{\$8,63}{\sqrt{200}} = \frac{\$8,63}{14,14213562} = \$0,610.$$

Таким образом, при использовании среднего значения выборки, равного \$19,42, в качестве оценки неизвестного значения μ для всех покупателей ошибка составила приблизительно \$0,610. Следует отметить, насколько меньше величина стандартной ошибки (\$0,610) по сравнению со стандартным отклонением (\$8,63).

Если опрошен только один покупатель и (по незнанию) попытались использовать его ответ в качестве оценки планируемых покупок для всех покупателей, то ошибка составит приблизительно \$8,63. Увеличив объем выборки до 200 и используя выборочное среднее, вы значительно уменьшите ошибку — приблизительно до \$0,610.

Поправка для малой генеральной совокупности

Когда объем генеральной совокупности настолько мал, что выборка составляет большую часть совокупности, стандартную ошибку можно уменьшить, введя в формулу **поправку для конечной генеральной совокупности** $\sqrt{(N-n)/N}$, чтобы получить **скорректированную стандартную ошибку**.

Если размер выборки практически равен генеральной совокупности, информация о совокупности является достаточно полной. Действительно, когда объем выборки равен объему совокупности (т.е. $N = n$), информация является абсолютно полной и стандартная ошибка должна равняться нулю. Для корректировки стандартной ошибки с целью получения большей точности используют следующую формулу[10].

[10] Теоретическое обоснование этой формулы можно найти, например, в работах W. G. Cochran, *Sampling Techniques*, 3rd ed. New York: Wiley, 1977, Equation 2.20, p.26, L. Kish, *Survey Sampling* New York: Wiley, 1967, Equation 2.2.2, p.41.

Скорректированная стандартная ошибка

(Поправочный коэффициент для конечной совокупности) × (Стандартная ошибка) =

$$= \sqrt{\frac{N-n}{N}} \times S_{\bar{x}} = \sqrt{\frac{N-n}{N}} \times \frac{S}{\sqrt{n}}$$

Если размер выборки приближается к размеру генеральной совокупности, значение $N-n$ уменьшается и значение скорректированной стандартной ошибки также уменьшается, что отражает высокое качество этой почти полной выборки. Когда объем совокупности N большой, то величина корректирующего коэффициента приближается к 1 и он почти не влияет на значение стандартной ошибки.[11]

Может возникнуть вопрос, почему получается так, что в случае большой генеральной совокупности ее размер N не оказывает влияния на стандартную ошибку, которая зависит только от информации о выборке, т.е. от n и S. Происходит это потому, что стандартная ошибка отражает в первую очередь не какую-то определенную характеристику генеральной совокупности, а случайность *процесса выборки*. В случае небольшой выборки из крупной генеральной совокупности отдельные значения в выборке не могут сильно "взаимодействовать" друг с другом, поскольку в процессе построения выборки отсутствуют возвраты и свойства выборки (такие как изменчивость среднего) будут оставаться практически неизменными, даже если удвоить размер генеральной совокупности (сохранив неизменными остальные ее характеристики). С другой стороны, при небольшом размере генеральной совокупности отсутствие возвратов оказывает более сильное влияние, ограничивая возможность выбора, причем это влияние зависит от размера генеральной совокупности. На рис. 8.4.3 показана ситуация, при которой выборочные значения не зависят от размера (большого) генеральной совокупности.

Не всегда целесообразно вводить в вычисления поправку на конечность генеральной совокупности, даже если формально это можно делать. Иногда основа выборки, из которой вы собственно и делаете отбор, в действительности не является интересующей вас генеральной совокупностью. Если вы предполагаете, что ваша основа выборки представляет собой некоторую случайную выборку из намного большей по размеру генеральной совокупности, то в этом случае лучше не использовать поправочный коэффициент. **Теоретическую генеральную совокупность** можно определить как намного большую по размеру, иногда воображаемую, генеральную совокупность, которую представляет ваша выборка. Если вас интересует теоретическая генеральная совокупность, не используйте поправочный коэффициент. В то же время, если вы работаете с основой выборки и не выходите за ее пределы, то поправочный коэффициент окажется полезным, поскольку его использование уменьшит вариацию.

[11] Когда N большое и n составляет небольшую часть совокупности, корректирующий коэффициент для совокупности, имеющей небольшой размер, уменьшает стандартную ошибку примерно на половину этой части, т.е. на $n/(2N)$. Если, к примеру, выборка составляет 8% большой совокупности, то корректировка уменьшит стандартную ошибку примерно на 4%. (В данном случае точная коррекция составит 4,08%, что достаточно близко к 4%.)

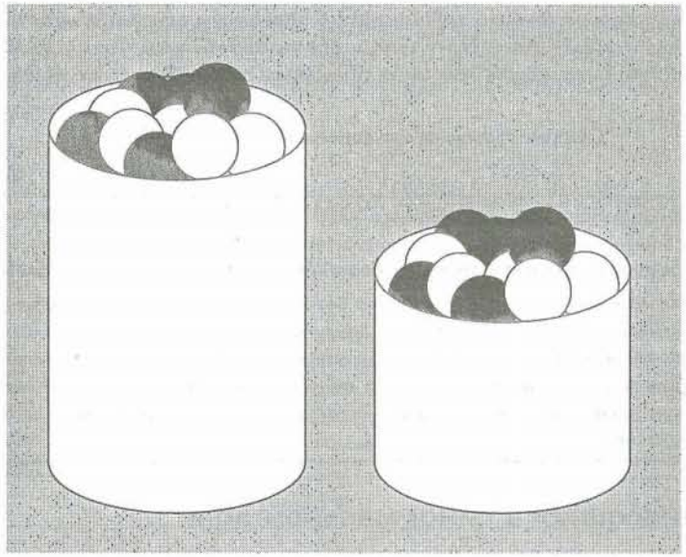


Рис. 8.4.3. При извлечении небольшой выборки из большой генеральной совокупности размер генеральной совокупности не влияет на стандартную ошибку. Имеются две отличающиеся размером, но одинаковые по содержанию урны. Если из каждой урны извлечь случайным образом по несколько шаров, распределение количества черных шаров в обеих выборках должно быть одинаковым. Каждая выборка дает информацию о процентном содержании черных шаров в соответствующей урне

Например, допустим, из списка 300 последних покупателей вы отобрали случайно 50, чтобы провести среди них опрос о качестве обслуживания. Если вас интересуют только эти 300 последних покупателей из списка, то можно спокойно уменьшать первоначально вычисленную приблизительно на 8,71%. Однако если вы хотите узнать о *покупателях вообще*, потенциально очень большой группе, которая представлена вашим списком из трехсот человек, то поправку вводить не следует. Если же вы ошибочно используете поправку, то вы можете ввести в заблуждение самого себя, считая, что полученные результаты являются более точными, чем они есть на самом деле.

При наличии сомнений лучше *отказаться* от поправочного коэффициента.

Пример. *Качество ежедневной продукции*

Из 48 грузовиков, загруженных произведенной на вашей фабрике продукцией, вы случайным образом отобрали 10 с целью тщательной проверки качества. Во время проверки каждой партии продукции присваивали оценки от 1 до 20, где 20 означало оценку "отлично", 1 — "очень плохо". Были выставлены следующие оценки: 19, 20, 20, 17, 20, 20, 15, 18, 20 и 15. Среднее дневной выборки составляет 18,4, а стандартное отклонение отдельных партий — 2,065591. Нескорректированная стандартная ошибка составляет 0,653.

Если вас интересует, насколько значение среднего дневной выборки (18,4) отличается от значения дневного среднего всех 48 партий (которое неизвестно, так как для проверки взяты только 10 партий), то

можно использовать корректирующий коэффициент. В этом случае действительно следует использовать этот коэффициент, поскольку для выборки была отобрана значительная часть (10/48 = 20,8%) генеральной совокупности. Тогда скорректированная стандартная ошибка будет следующей:

$$\text{Скорректированная стандартная ошибка} = \sqrt{\frac{N-n}{N}} \times S_{\bar{x}} =$$

$$= \sqrt{\frac{48-10}{48}} \times \frac{2{,}065591}{\sqrt{10}} = 0{,}889757 \times 0{,}653197 = 0{,}581.$$

Корректировка уменьшила исходную стандартную ошибку на 11% (= 1 – 0,889757), с 0,653 до 0,581.

С другой стороны, если вас интересует, насколько значение средней оценки качества дневной выборки отличается от значения оценки *качества вообще* продукции всей фабрики, не стоит вводить поправку, а следует использовать нескорректированную стандартную ошибку (0,653). По существу, вы будете пытаться распространить полученные результаты на очень большую *теоретическую генеральную совокупность*, включающую все партии товара, которые *могли бы* быть произведены сегодня на этой фабрике при существующих условиях.

Стандартная ошибка биномиальной доли

В случае биномиального распределения имеют место две ошибки: одна для частоты X, другая для доли ρ. Стандартная ошибка S_x показывает неопределенность, или изменчивость, наблюдаемой частоты, и ее легко вычислить исходя из информации в выборке. Аналогично этому стандартная ошибка S_ρ показывает неопределенность в наблюдаемой доле. Для вычисления этих двух ошибок можно использовать те же формулы, что и для вычисления стандартных отклонений (генеральной совокупности) σ_x и σ_ρ в главе 7, заменив только неизвестную долю в генеральной совокупности π на ее выборочную оценку ρ. Это общий подход: использовать наилучшую доступную информацию из выборки (в данном случае ρ) вместо информации о генеральной совокупности (например, π), которая необходима, но недоступна. Ниже приведены формулы для вычисления стандартных отклонений генеральной совокупности и стандартных ошибок (оцененных из выборочных данных) для биномиального распределения[12].

Биномиальное стандартное отклонение и стандартная ошибка

	Частота событий, X	Доля или процент, $\rho = X/n$
Стандартное отклонение (для генеральной совокупности)	$\sigma_x = \sqrt{n\pi(1-\pi)}$	$\sigma_\rho = \sqrt{\frac{\pi(1-\pi)}{n}}$
Стандартная ошибка (оцененная по выборке)	$S_x = \sqrt{n\rho(1-\rho)}$	$S_\rho = \sqrt{\frac{\rho(1-\rho)}{n}}$

[12] Следует отметить, что данные формулы позволяют вычислять стандартные ошибки для биноминального распределения непосредственно. Иными словами, не нужно вначале вычислять стандартное отклонение S, а затем делить его на квадратный корень из n, что приходится делать в случае работы со списком чисел (не биноминальная ситуация).

Например, если обнаружено, что 8 станков из 50 являются бракованными, то наблюдаемая биномиальная доля будет равна 0,16, или 16%, с неопределенностью $S_\rho = 0{,}0518$, или 5,18 процентных единиц. Наблюдаемая частота X равна 8 с неопределенностью $S_x = 2{,}59$.

Пример. *Опрос покупателей*

Вы опросили 937 человек, из которых 302, или 32,2%, решили приобрести ваше изделие. Вас интересует, насколько надежны эти цифры. В частности, насколько они отличаются от значений для всей значительно большей генеральной совокупности? Для ответа на этот вопрос необходимо определить стандартную ошибку.

Предположим, что имеет место биномиальное распределение, так как людей из генеральной совокупности отбирали независимо и случайным путем. Итак, вы знаете, что $n = 937$, $X = 302$ и $\rho = 0{,}322$, или 32,2%. Однако неизвестно π — интересующий нас и в то же время неизвестный процент во всей генеральной совокупности. Можно использовать стандартную ошибку (оценку стандартного отклонения), так как у вас есть *оценка* π, а именно наблюдаемое значение 0,322.

	Число людей, X	Доля или процент $\rho = X/n$
Стандартная ошибка (для биномиального распределения)	$S_x = \sqrt{n\rho(1-\rho)} = \sqrt{937 \times 0{,}322(1-0{,}322)} =$ 14,3 человек	$S_\rho = \sqrt{\frac{\rho(1-\rho)}{n}} = \sqrt{\frac{0{,}322(1-0{,}322)}{937}} =$ 0,0153 или 1,53%

Наблюдаемое значение 302 человека приблизительно на 14,3 человека (в большую или меньшую сторону) отличается от неизвестного значения, которое вы могли бы ожидать получить в подобного рода исследованиях такой генеральной совокупности. Наблюдаемая доля людей, равная 32,2%, примерно на 1,53 процентных единицы отличается от действительной неизвестной процентной доли во всей генеральной совокупности.

8.5. Другие методы построения выборки

Случайная выборка — не единственный метод построения (извлечения) выборки из генеральной совокупности. Существует много других методов, каждый из которых имеет свои преимущества и недостатки. Некоторые, как, например, стратифицированная случайная выборка, тщательно используют принципы случайного отбора. Другие, например систематическая выборка, основаны на существенно других подходах и являются слабой базой для статистического анализа.

При выборе метода необходимо взвесить и сопоставить негативную критику, которой могут быть подвергнуты ваши результаты, и стоимость получения данных. Для внутреннего исследования в дружеской рабочей обстановке, без сложностей во взаимоотношениях внутри организации (если такие места еще существуют вообще!), нет необходимости в тщательном построении случайной выборки. Однако для внешнего исследования, которое будет использовано нейтральной, а может быть, и враждебно настроенной стороной, например подавшей иск, когда представители другой стороны могут подвергнуть сомнению ваш профессионализм, имеет смысл обратить внимание на детали и использовать тщательно спланированные методы построения случайной выборки.

Стратифицированная случайная выборка

Иногда генеральная совокупность содержит ясные, известные, легко идентифицируемые группы. Если вы строите случайную выборку из всей генеральной совокупности, каждый такой сегмент, или *слой (страта)*, может быть недостаточно или, наоборот, избыточно представлен в этой выборке по сравнению с тем, как он представлен в генеральной совокупности.[13] Это может привнести в результаты дополнительную случайность, поскольку известная информация об этих группах не будет использована.

Стратифицированную случайную выборку получают путем осуществления случайной выборки отдельно в каждой страте (сегменте, или слое) генеральной совокупности. Если генеральная совокупность однородна (гомогенна) внутри каждой страты, но отдельные страты заметно отличаются друг от друга, стратификация может увеличить точность статистического анализа. Стратификация также облегчает управление исследованием, поскольку появляется возможность поручить отбор определенным филиалам центрального офиса.

Размеры выборки для каждой из страт могут быть разными. Не обязательно отбирать одинаковое количество элементов из каждой страты или планировать размер выборки для страты в соответствии с ее процентным содержанием в генеральной совокупности. Это позволяет определять размеры выборок для страт исходя из затрат и ресурсов. Для одних страт процесс отбора может быть сложнее и дороже, чем для других, и для этих страт вы будете стараться использовать меньшие по размеру выборки. Другие страты могут иметь большую изменчивость, и поэтому для них вы будете использовать большие по размеру выборки.

В табл. 8.5.1 содержатся обозначения для размеров генеральной совокупности, размеров выборок, выборочных средних и выборочных стандартных отклонений для каждой из страт.

Осталось показать, как из оценок, вычисленных для каждой страты, получить оценку для всей генеральной совокупности и каким образом определить стандартную ошибку для этой результирующей оценки. Для каждой страты строится случайная выборка и вычисляется выборочное среднее. Оценку среднего генеральной совокупности на основе стратифицированной выборки вычисляют как взвешенное среднее выборочных средних отдельных страт. При этом веса вычисляют исходя из размеров страт. Таким образом, более крупные страты оказывают большее влияние на результат. Сначала каждое выборочное среднее необходимо умножить на размер соответствующей страты, затем сложить полученные результаты и разделить на общий размер генеральной совокупности.

Вычисление среднего для стратифицированной выборки

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + \ldots + N_L\bar{X}_L}{N_1 + N_2 + \ldots + N_L} = \frac{\sum_{i=1}^{L} N_i\bar{X}_i}{N},$$

где N — размер всей генеральной совокупности $N_1 + N_2 + \ldots + N_L$.

[13] Для обозначения сегмента или стоя генеральной совокупности в английском языке используется слово stratum (в единственном числе) или strata (во множественном числе).

Таблица 8.5.1. Обозначения для стратифицированной выборки

Страта	Размер совокупности	Размер выборки	Среднее выборки	Стандартное отклонение выборки
1	N_1	n_1	$\bar{X}_1$	S_1
2	N_2	n_2	$\bar{X}_2$	S_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
L	N_L	n_L	$\bar{X}_L$	S_L

Какова величина изменчивости результирующей оценки? Как всегда, ответ на этот вопрос дает стандартная ошибка. Она вычисляется следующим образом, исходя из стандартных отклонений для отдельных страт. Для каждой страты квадрат соответствующего стандартного отклонения умножают на квадрат размера страты и делят на размер выборки из этой страты. Затем все полученные произведения складывают, извлекают корень квадратный и делят на размер всей генеральной совокупности.

Стандартная ошибка для стратифицированной выборки

$$S_{\bar{X}} = \frac{1}{N}\sqrt{\frac{N_1^2S_1^2}{n_1} + \frac{N_2^2S_2^2}{n_2} + \ldots + \frac{N_L^2S_L^2}{n_L}} = \frac{1}{N}\sqrt{\sum_{i=1}^{L}\frac{N_i^2S_i^2}{n_i}}.$$

Если размеры некоторой страты настолько малы, что выборка составляет значительную часть этой страты, то можно использовать поправочный коэффициент, чтобы получить скорректированное (более точное) значение стандартной ошибки. Для каждой страты ее размер умножают на квадрат стандартного отклонения и на разность между размерами страты и выборки, а затем делят на размер выборки. Полученные результаты складывают, извлекают корень квадратный и делят на размер всей генеральной совокупности.

Скорректированная стандартная ошибка для стратифицированной выборки

$$\text{Скорректированная стандартная ошибка} = \frac{1}{N}\sqrt{\frac{N_1(N_1-n_1)S_1^2}{n_1} + \frac{N_2(N_2-n_2)S_2^2}{n_2} + \ldots + \frac{N_L(N_L-n_L)S_L^2}{n_L}} = \frac{1}{N}\sqrt{\sum_{i=1}^{L}\frac{N_i(N_i-n_i)S_i^2}{n_i}}$$

Пример. *Поправка на осведомленность потребителя*

Для разработки маркетинговой стратегии продвижения высокотехнологичной аудио- и видеопродукции вам требуется соответствующая информация о потенциальных покупателях. В зависимости от осведомленности о данной технологии покупателей можно достаточно естественным образом разделить на две группы. Группа осведомленных покупателей желает знать технические особенности продукции; группе неподготовленных покупателей необходима лишь базовая информация общего характера.

Чтобы определить, сколько денег в этом году планирует потратить на вашу продукцию типичный потенциальный покупатель, вы решили использовать в своем исследовании стратифицированную случайную выборку. Это разумно, так как вы ожидаете, что группа осведомленных покупателей планирует более крупные расходы. Стратификация позволит уменьшить общую вариацию возможных объемов затрат на вашу продукцию.

Основа вашей выборки — это список имен и адресов 14 000 потенциальных покупателей, полученный из маркетинговой фирмы. Покупатели в списке уже классифицированы: 2532 из них являются осведомленными и 11468 — неподготовленными. Вы решили отобрать 200 осведомленных и 100 неподготовленных покупателей для детального опроса, сделав акцент на сегменте подготовленных покупателей, поскольку их ожидаемая покупательская способность выше. Ниже приведены результаты.

Страта	Объем совокупности	Объем выборки	Среднее выборки	Стандартное отклонение выборки
Неподготовленные	$N_1 = 11468$	$n_1 = 100$	$\bar{X}_1 = \$287$	$S_1 = \$83$
Осведомленные	$N_2 = 2532$	$n_2 = 200$	$\bar{X}_2 = \$1253$	$S_2 = \$454$

Полученные для двух слоев оценки очень интересны. Результаты подтвердили ваши предположения о том, что осведомленные потенциальные покупатели планируют потратить больше денег!

Чтобы найти среднее значение расходов одного потенциального покупателя для всей генеральной совокупности, вычислим взвешенное среднее:

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2} = \frac{11468 \times 287 + 2532 \times 1253}{11468 + 2532} = \frac{6463912}{14000} = \$462.$$

Величина результирующего среднего, $462, намного ближе к величине расходов на покупки у неподготовленных покупателей ($287), чем к величине расходов у осведомленных ($1253). Это следствие того, что страта неподготовленных покупателей составляет значительно большую часть генеральной совокупности, чем страта осведомленных. Если даже увеличить объем выборки для осведомленных покупателей в два раза, это просто улучшит информацию об этих покупателях, но не усилит влияние этого слоя генеральной совокупности на окончательный результат.

Какова неопределенность полученной оценки расходов в объеме $462 на одного покупателя? Стандартная ошибка будет следующей:

$$S_{\bar{x}} = \frac{1}{N}\sqrt{\frac{N_1^2 S_1^2}{n_1} + \frac{N_2^2 S_2^2}{n_2}} = \frac{1}{14000}\sqrt{\frac{11468^2 \times 83^2}{100} + \frac{2532^2 \times 454^2}{200}} = \frac{1}{14000}\sqrt{9060070003 + 6607073115} = \$8{,}94.$$

Как можно настолько точно, с такой небольшой стандартной ошибкой $8,97, определить средние затраты одного покупателя, в то время как разброс расходов отдельных покупателей для разных групп равен $83 или $454? Ответ кроется в том, что оценивается среднее значение расходов, а не расходы отдельных покупателей. И если принять во внимание только сегмент неподготовленных покупателей, который составляет большую часть рынка, стандартная ошибка составит $83/10 или только чуть больше $8.

Каковы в данном случае преимущества стратифицированной выборки по сравнению с обычной случайной выборкой размером 300 из всей генеральной совокупности? Стратифицированная выборка дает больше возможностей контролировать вариацию. Вместо объединения больших (для подготовленных покупателей) и небольших (для неосведомленных) расходов в одну выборку с большой вариацией вы разделили эту вариацию в соответствии с ее источниками, которые вам известны. Вследствие этого вы получили более точный результат. Можно показать (подробности здесь не приводятся), что без стратификации стандартная ошибка была бы в три раза выше, чем полученное значение $8,94. И помните, что трехкратное сокращение стандартной ошибки обычно достигается только при девятикратном увеличении размера выборки (поскольку $9=3^2$). Используя стратификацию, вы с помощью выборки объемом 300 достигли результатов, сравнимых с простой случайной выборкой размером примерно 900. Таким образом, стратификация может быть отличным способом сокращения расходов на исследование!

При решении различных задач стратификация может быть в большей или в меньшей степени полезной, чем в рассмотренном нами примере. Стратификация более полезна тогда, когда в пределах каждого слоя (страты) элементы подобны, а сами слои отличаются один от другого. Иными словами, слои выделяют в генеральной совокупности содержательно значимые и важные части.

Если вычислить стандартную ошибку более тщательно, используя поправочный коэффициент, то окажется, что стандартная ошибка уменьшается с $8,94 до $8,77. Будем считать, что вы решили оставить значение без поправки, 8,94, поскольку не хотите преувеличивать точность результатов и, кроме того, заинтересованы в возможном распространении результатов, полученных из списка 14 000 покупателей, на значительно большую, чем представленная этой основой выборки, теоретическую генеральную совокупность.

Пример. *Цена обычного костюма в универмаге*

Рассмотрим магазин с двумя отделами: обычной одежды (где продают обычные костюмы) и модной одежды (где продают высококлассные дорогие костюмы). Отдел обычной одежды имеет больший общий объем продаж, но более низкую стоимость одного костюма. Отдел модной одежды имеет меньшее количество покупателей, но цена одного костюма выше. Чтобы определить суммарный объем продаж в универмаге, руководству хотелось бы иметь единую цифру, которая выражала бы цену, которую обычно выручают от продажи костюма.

На рис. 8.5.1 показана типичная ситуация: 90% костюмов продают в отделе обычной одежды (где стоимость одного костюма $60) и 10% продают в отделе высокой моды (где один костюм стоит $450). В верхней части рисунка показана репрезентативная выборка из 10 покупателей и средняя цена костюма в $99 (поскольку было куплено 9 костюмов по $60 и один костюм, стоимостью $450).

В нижней части рис. 8.5.1 показана нерепрезентативная выборка, включающая по одному покупателю из каждого отдела. Если в качестве типичной стоимости костюма просто вычислить обычное среднее арифметическое стоимости двух костюмов, один из которых стоит $60, а второй — $450, то получим значение $255, что, понятно, неверно. Дело в том, что один покупатель из отдела общих продаж представляет значительно большую часть генеральной совокупности, чем один покупатель из отдела высокой моды.

Использование стратифицированной выборки позволяет исправить эту проблему путем применения формулы для вычисления взвешенного среднего. Присвоив вес 90% покупателю из отдела общих про-

даж, а вес 10% — покупателю из отдела высокой моды и вычислив затем взвешенное среднее, мы получим действительно верный ответ[14]:

$$\text{Среднее взвешенное} = 0{,}90 \times \$60 + 0{,}10 \times \$450 = \$99 .$$

Таким образом, становится ясно, почему среднее, вычисленное на данных стратифицированной выборки, является верным. Использованные при вычислении взвешенного среднего веса отражают значимость каждого слоя выборки в генеральной совокупности.

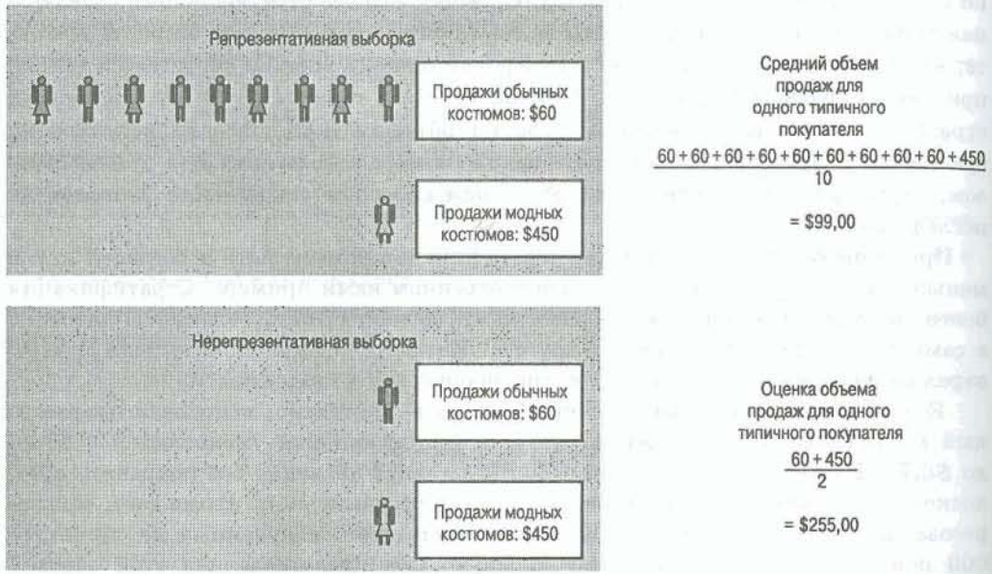


Рис. 8.5.1. Методы тщательной стратификации выборки могут корректировать проблемы нерепрезентативного отбора. Простое среднее для нерепрезентативной выборки, равное $255, является неверным. Однако взвешенное среднее для этой же выборки даст верный результат — $99

Систематическая выборка

Для получения **систематической выборки** необходимо выбрать в основе выборки одну случайную начальную точку и затем производить отбор элементов в основе выборки с некоторым постоянным шагом. Можно легко построить такую выборку, если, скажем, отбирать каждый пятый элемент из основы выборки, как это показано на рис. 8.5.2. Можно также случайным образом выбрать начальную точку отбора и таким образом привнести в выборку элемент случайности. Но все равно такой метод систематического отбора имеет серьезные проблемы из-за невозможности оценить точность получаемой выборки.

Если вы хотите построить систематическую выборку объемом n из генеральной совокупности объемом N, то интервал между выбираемыми элементами будет равен N/n.[15] Если в качестве начальной точки отбора вы выбрали число между 1 и N/n, то выборочное среднее будет обоснованной оценкой среднего генеральной совокупности в том смысле, что оно будет *несмещенным*, т.е. его величина не будет систематически завышенной или заниженной. И это хорошо.

[14] О взвешенном среднем идет речь в главе 4.

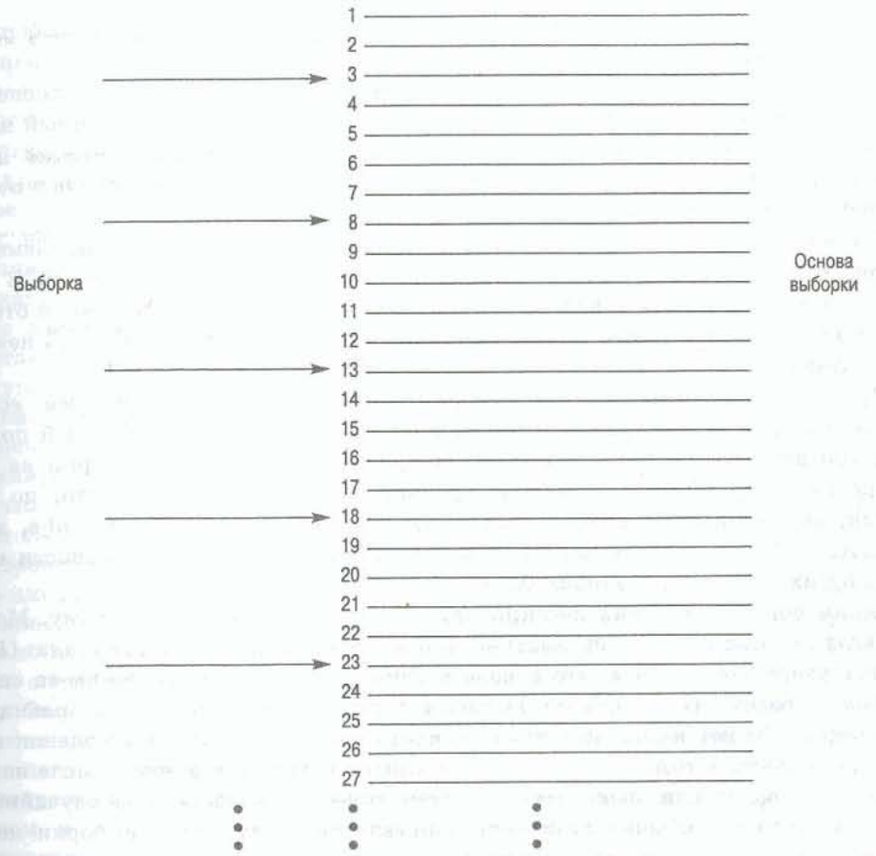


Рис. 8.5.2. Систематическую выборку строят путем целенаправленного отбора данных в генеральной совокупности. В данном случае из основы выборки отбирается каждый пятый элемент начиная с элемента с номером три

Плохо то, что невозможно узнать, *насколько хороша* эта оценка. Если задать вопрос: "Какова стандартная ошибка?", то ответом будет: "А кто знает? Выборка не является действительно случайной". По словам Эдварда Деминга (W. Edward Deming) (помимо всего прочего, знаменитого тем, что он привнес качество в японские товары), "один из методов проведения выборки, который я, так же как и другие специалисты, использовал, заключался в отборе каждого k-го элемента

[15] Если N/n не является целым числом, то появляются некоторые небольшие технические трудности, требующие специального внимания. Более подробное обсуждение этого и других аспектов систематического отбора см. в главе 4 книги Kish, *Survey Sampling*.

начиная со случайно выбранной начальной точки (систематическая выборка)... Поскольку нет повторений (имеется в виду повторение случайного отбора. — *Прим. ред.*), то нет и возможности вычислить несмещенную оценку дисперсии оценок, полученных с применением этой процедуры... Метод, основанный на повторении (случайная выборка), настолько прост в применении, что не имеет смысла использовать такие методы построения оценок, которые не дают ответы на необходимые вопросы".[16]

Систематическая выборка может оказаться неудачной, если список элементов упорядочен некоторым, важным с определенной точки зрения образом. В этой ситуации оценка может определяться одной случайной начальной точкой отбора, и, например, малый номер начальной точки может сразу гарантировать невысокую оценку.

Применение систематической выборки обернется серьезной неудачей, если в основе выборки существует некоторый повторяемый фрагмент, который по размеру соответствует интервалу отбора. Например, если во время сборки на конвейере каждому 50-му автомобилю уделяют особое внимание и если, по воле случая, вы отбираете в свою систематическую выборку именно эти 50-е, ваши результаты будут полностью бесполезны в отношении репрезентативности качества других *обычных* автомобилей.

Таким образом, систематический отбор можно оценивать по-разному. Можно оправдать использование систематической выборки в следующих случаях: (1) если есть уверенность в том, что основа выборки не упорядочена каким-то специальным образом, (2) в основе выборки отсутствуют важные повторяющиеся фрагменты, (3) нет необходимости оценивать качество полученных оценок и (4) есть уверенность в том, что никто не усомнится в вашем здравом смысле исходя из того, что вы отдали предпочтение систематической выборке, а не случайной.

Ввиду того что обычно стоимость действительно случайной выборки не намного превышает стоимость систематической выборки, можно только удивляться, почему до сих пор в некоторых областях бизнеса все еще применяют систематическую выборку. Я и удивляюсь этому.

8.6. Дополнительный материал

Резюме

Выборку используют для изучения системы, которая является настолько большой, что ее полное исследование стоит слишком дорого. **Генеральная совокупность** — это набор элементов (люди, объекты и т.п.), которые необходимо изучить. **Выборка** — это меньший набор элементов, извлеченных из генеральной совокупности. Выборку называют **репрезентативной**, если каждое свойство (или комбинация свойств) наблюдается в выборке с той же частотой, что и в генеральной совокупности. О выборке, которая не является репрезентативной, говорят, что она имеет **смещение. Основа выборки** позволяет по числу из интервала от 1 до N (размер совокупности) получить доступ к элементу генеральной совокупности. Говорят, что выборку извлекают **без возврата**, если никакой элемент генеральной совокупности не может быть отобраны в выборку более одного раза. Говорят, что выборку извлекают **с возвратом**, если элемент генеральной совокупности может быть отобран в выборку более одного раза. Выборку, которая включает всю совокупность ($n = N$), называют **переписью.**

Статистикой, или выборочной статистикой, называют любое число, вычисленное на данных выборки. **Параметром, или параметром генеральной совокупности,** называют любое число, рассчитанное для всей генеральной совокупности. **Оценочная функция (оценка)** — это выборочная статистика, которая используется как предполагаемое значение параметра генеральной совокупности. Фактическое значение, вычисленное на данных, называют **оценкой. Ошибка оценки** представляет разность между оценочной функцией (или оценкой) и параметром совокупности. Ошибка оценки обычно неизвестна. Оценка называется **несмещенной,** если она не является систематически завышенной или заниженной по отношению к значению соответствующего параметра генеральной совокупности. **Случайную выборку,** или **простую случайную выборку,** извлекают таким образом, что: (1) все элементы генеральной совокупности имеют *одинаковые вероятности быть отобранными* и (2) элементы генеральной совокупности *отбираются независимо* друг от друга. **Таблица случайных чисел** — это последовательность цифр, в которой все цифры от 0 до 9 появляются независимо друг от друга и с одинаковой вероятностью 1/10. Использование такой таблицы для последовательного отбора различных элементов генеральной совокупности служит одним из способов извлечения случайной выборки без возврата.

Пробное исследование — это маломасштабная версия исследования, разработанная и выполненная, чтобы помочь найти и решить проблемы до проведения настоящего полного исследования.

Центральная предельная теорема устанавливает, что для случайной выборки объемом n наблюдений из генеральной совокупности справедливы следующие утверждения.

1. С ростом n распределение как *среднего*, так и *суммы* все более приближается к нормальному.
2. Средние значения и стандартные отклонения распределений среднего и сумм вычисляют по приведенным ниже формулам, где μ — среднее значение и σ — стандартное отклонение значений элементов генеральной совокупности.

	Случайная величина	
	среднее	общая сумма
Среднее	$\mu_{\bar{x}} = \mu$	$\mu_{сумм} = n\mu$
Стандартное отклонение	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$\sigma_{сумм} = \sigma\sqrt{n}$

Применяя центральную предельную теорему, можно найти вероятности для суммы или среднего случайной выборки в стандартной таблице нормального распределения вероятностей, предварительно вычислив соответствующие среднее и стандартное отклонения по формулам, приведенным выше.

[16] W. Edward Deming, *Sample Design in Business Research*. New York: Wiley, 1960, p.98.

Всякая статистика, вычисленная на данных случайной выборки, характеризуется вероятностным распределением, которое называют **выборочным распределением** этой статистики. **Стандартная ошибка** статистики, т.е. оценка стандартного отклонения ее выборочного распределения, приближенно показывает, насколько значение статистики может отличаться от своего среднего значения (параметра генеральной совокупности). **Стандартная ошибка среднего** (или, кратко, просто **стандартная ошибка**) приближенно показывает, насколько среднее выборки $\overline{X}$ (случайная наблюдаемая величина) отличается от среднего генеральной совокупности μ (фиксированная неизвестная величина):

$$\text{Стандартная ошибка} = S_{\overline{x}} = \frac{S}{\sqrt{n}}.$$

Стандартная ошибка уменьшается с увеличением размера выборки n (при прочих равных условиях), отражая тот факт, что большая по размеру выборка содержит больше информации и таким образом достигается большая точность.

Когда объем генеральной совокупности настолько мал, что выборка составляет достаточно большую часть генеральной совокупности, стандартную ошибку можно уменьшить, введя в формулу **корректирующий** (поправочный) **коэффициент для конечной совокупности**, чтобы получить **уточненную (откорректированную) стандартную ошибку**:

$$\text{(Корректирующий коэффициент для конечной совокупности)} \times$$

$$\times\text{(Стандартная ошибка)} = \sqrt{\frac{N-n}{N}} \times S_{\overline{x}} = \sqrt{\frac{N-n}{N}} \times \frac{S}{\sqrt{n}}.$$

Теоретическую (идеальную) генеральную совокупность можно определить как очень большую, иногда предполагаемую (воображаемую) генеральную совокупность, которую представляет ваша выборка. Если вас интересует теоретическая генеральная совокупность, не используйте поправку на конечность генеральной совокупности. С другой стороны, если необходимо сделать вывод об основе выборки, не выходя за ее пределы, то поправка может быть полезной, так как ее использование уменьшает вариацию системы. Если есть сомнения, лучше *не* использовать поправку.

Для биномиального распределения стандартные отклонения (генеральной совокупности) и стандартные ошибки (выборки) как для X (для частоты), так и для $p = X/n$ (для доли) вычисляют следующим образом.

	Биномиальная частота событий, X	**Биномиальная доля, или процент, $p = X/n$**
Стандартное отклонение (для генеральной совокупности)	$\sigma_x = \sqrt{n\pi(1-\pi)}$	$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$
Стандартная ошибка (оценка для выборки)	$S_x = \sqrt{np(1-p)}$	$S_p = \sqrt{\frac{p(1-p)}{n}}$

Стандартная ошибка S_p показывает неопределенность, или изменчивость, в наблюдаемой доле p, а стандартная ошибка S_X — неопределенность в наблюдаемой частоте X.

Стратифицированную случайную выборку получают путем извлечения случайной выборки отдельно из каждой страты (слоя, типической группы или сегмента) генеральной совокупности. Если генеральная совокупность однородна внутри каждой страты, но страты заметно отличаются друг от друга, стратификация может увеличить точность статистического анализа. Для совокупности с L стратами и количеством элементов N_i в i-й страте используют такие обозначения: размер выборки — n_i, выборочное среднее — $\overline{X}_i$ и выборочное стандартное отклонение — S_i. Чтобы из этих средних получить одно число, которое характеризует всю генеральную совокупность, вычисляют взвешенное среднее. Ниже приведены формулы для вычисления взвешенного среднего и его стандартной ошибки:

$$\overline{\overline{X}} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2 + \ldots + N_L\overline{X}_L}{N_1 + N_2 + \ldots + N_L} = \frac{\sum_{i=1}^{L} N_i\overline{X}_i}{N};$$

$$S_{\overline{\overline{x}}} = \frac{1}{N}\sqrt{\frac{N_1^2S_1^2}{n_1} + \frac{N_2^2S_2^2}{n_2} + \ldots + \frac{N_L^2S_L^2}{n_L}} = \frac{1}{N}\sqrt{\sum_{i=1}^{L}\frac{N_i^2S_i^2}{n_i}};$$

Скорректированная стандартная ошибка =

$$= \frac{1}{N}\sqrt{\frac{N_1(N_1-n_1)S_1^2}{n_1} + \frac{N_2(N_2-n_2)S_2^2}{n_2} + \ldots + \frac{N_L(N_L-n_L)S_L^2}{n_L}} =$$

$$= \frac{1}{N}\sqrt{\sum_{i=1}^{L}\frac{N_i(N_i-n_i)S_i^2}{n_i}}.$$

Используйте скорректированную стандартную ошибку как поправку на конечность совокупности, когда для какой-либо из страт соответствующая выборка составляет значительную часть этой страты.

Систематическую выборку получают, выбирая в основе выборки случайную начальную точку и затем отбирая элементы основы выборки начиная с этой точки через постоянный интервал (с постоянным шагом отбора). Хотя среднее систематической выборки является несмещенной оценкой среднего генеральной совокупности (т.е. не является постоянно завышенным или заниженным), применение такого метода связано с определенными серьезными проблемами. Невозможно определить, насколько удовлетворительной является оценка, так как для нее нет надежной стандартной ошибки. Особенно серьезные проблемы могут возникнуть, если элементы генеральной совокупности упорядочены в основе выборки особым образом или если в основе выборки есть повторяющиеся группы элементов. В связи с тем что построение случайной выборки обходится, как правило, не намного дороже, чем построение систематической выборки, использования систематической выборки желательно избегать.

Основные термины

- Генеральная совокупность (population), 341
- Выборка (sample), 341
- Репрезентативная выборка (representative), 343
- Смещение (bias), 343
- Основа генеральной совокупности (frame), 343
- Выборка без возврата (sampling without replacement), 344
- Выборка с возвратом (sampling with replacement), 344
- Перепись (census), 345
- Параметр выборки (sample statistic), или статистика, 345
- Параметр генеральной совокупности (population parameter), 345
- Оценочная функция (estimator), 345
- Оценка (estimation), 345
- Ошибка оценки (error of estimation), 345
- Несмещенная оценка (unbiased estimator), 345
- Случайная выборка или простая случайная выборка (random sample or simple random sample), 346
- Таблица случайных чисел (table of random digits), 347
- Пробное исследование (pilot study), 352
- Выборочное распределение (sampling distribution), 352
- Центральная предельная теорема (central limit theorem), 354
- Стандартная ошибка статистики (standard error of statistic), 359
- Стандартная ошибка среднего (standard error of the average), 360
- Коэффициент поправки на конечность генеральной совокупности (finite-population correction factor), 363
- Скорректированная стандартная ошибка (adjusted standard error), 363
- Теоретическая (идеализированная) генеральная совокупность (idealized population), 364
- Стратифицированная случайная выборка (stratified random sample), 368
- Систематическая выборка (systematic sample), 372

Контрольные вопросы

1. а) Что такое генеральная совокупность?

   б) Что такое выборка? В чем польза использования выборок?

   в) Что такое перепись? Всегда ли при наличии возможности нужно производить перепись?
2. а) Что такое репрезентативная выборка?

   б) Что такое смещенная выборка?

   в) Как можно извлечь репрезентативную выборку?
3. Что такое основа выборки? Какова ее роль при построении выборки?
4. а) Что такое случайная выборка?

   б) Почему случайная выборка является приближенно репрезентативной?

   в) Каковы отличия между случайной выборкой *с* возвратом и случайной выборкой *без* возврата?

   г) Что такое таблица случайных чисел? Каким образом ее используют при построении выборки?

   д) Как можно использовать электронные таблицы для построения случайной выборки?
5. а) Что такое пробное исследование?

   б) Какие проблемы могут возникнуть, если не провести пробное исследование?
6. а) Что такое статистика?

   б) Что такое параметр?
7. а) Что такое оценочная функция?

   б) Что такое оценка?

   в) Стандартное отклонение выборки равно 13,8. Является ли это число оценочной функцией или оценкой стандартного отклонения генеральной совокупности?

   г) Что такое ошибка оценки? Когда вы оцениваете неизвестное число, знаете вы величину этой ошибки или нет?
8. а) Что такое выборочное распределение статистики?

   б) Что такое стандартное отклонение статистики?
9. а) В чем суть центральной предельной теоремы?

   б) Утверждает ли центральная предельная теорема, что результаты отдельных наблюдений подчиняются нормальному распределению?

   в) Как вы можете пояснить утверждение, что среднее имеет нормальное распределение?

   г) Что такое среднее суммы независимых наблюдений случайной переменной? Какое стандартное отклонение имеет эта величина?

   д) Что такое среднее значение среднего независимых наблюдений случайной переменной? Какое стандартное отклонение имеет эта величина?
10. а) Что такое стандартная ошибка статистики?

    б) Каким образом стандартная ошибка характеризует качество информации, полученной в результате оценивания?

    в) Какова тенденция изменения стандартной ошибка при увеличении объема выборки n ?
11. а) Что такое поправка на конечность генеральной совокупности?

    б) Что такое уточненная стандартная ошибка?

в) Что такое теоретическая (идеализированная) совокупность?

г) В каком случае ваши результаты более ограничены: при использовании поправки на конечность генеральной совокупности или без ее использования?

12. Что характеризуют стандартные ошибки S_x и S_p для биномиального распределения?

13. а) Что такое стратифицированная случайная выборка?

б) В чем преимущества стратификации?

в) Когда стратификация является наиболее эффективной?

14. а) Что такое систематическая выборка?

б) Какие основные проблемы возникают при использовании систематической выборки?

в) Почему не существует надежной стандартной ошибки для выборочного среднего систематической выборки?

Задачи

1. На вашей фабрике по выпуску автоматических трансмиссий возникли определенные проблемы с качеством. Для проведения тщательного анализа вы приняли решение собрать информацию о завтрашней продукции. Для каждого из указанных ниже методов извлечения выборки дайте оценку с точки зрения того, является ли такая процедура хорошей, подходящей или необоснованной. Аргументируйте свой выбор.

а) Первые пять произведенных трансмиссий.

б) Восемнадцать трансмиссий, которые находятся за пределами фабрики по причине того, что они никогда не работали.

в) Каждая двадцатая произведенная трансмиссия.

г) Взятая в конце рабочего дня случайная выборка с использованием дневной продукции в качестве основы выборки.

д) Все явно бракованные трансмиссии плюс случайная выборка очевидно нормальных трансмиссий.

2. Какая из приведенных ниже выборок будет наиболее репрезентативной для совокупности всех зарегистрированных избирателей США?

а) Выборка из 200 человек из торгового района Денвера.

б) Выборка из 200 ваших друзей и друзей ваших друзей.

в) Выборка из 200 человек, взятая на основе случайных телефонных номеров.

г) Случайная выборка из 200 студентов университета штата Небраска.

3. Какая из приведенных ниже выборок будет наиболее репрезентативной для совокупности всех работающих в компании *IBM*?

а) Десять самых старых и опытных специалистов из исследовательского центра Thomas J. Watson Research Center.

б) Случайная выборка из 10 специалистов по ремонту компьютеров.

в) Десять сотрудников, отобранных как "наиболее типичные" сотрудники среднего звена управления.

г) Случайная выборка 10 работников из списка всех сотрудников фирмы *IBM*.

4. Рассмотрим опрос избирателей, спланированный таким образом, чтобы каждая семья имела равный шанс быть отобранной и в каждой отобранной семье был опрошен один зарегистрированный избиратель. Проанализируйте ситуацию, когда вероятность проголосовать за демократов выше в тех семьях, где есть один зарегистрированный избиратель, по сравнению с теми семьями, где есть более одного зарегистрированного избирателя. Особенно обратите внимание на то, будет ли "процент голосующих за демократов в данной выборке" несмещенной оценкой соответствующего процента для всех зарегистрированных избирателей? Если нет, то будет ли эта оценка существенно завышенной или, наоборот, существенно заниженной по сравнению с реальным процентом?

5. Из общего количества 684 супермаркетов, за которые вы отвечаете, вы извлекли выборку объемом 25. В ходе проверки этих 25 супермаркетов были зарегистрированы нарушения политики вашей компании. Определите, какие из указанных ниже величин являются статистиками, а какие параметрами?

а) Среднее количество нарушений в 25 проверенных супермаркетах.

б) Среднее количество нарушений, которые могли бы быть зафиксированы при проверке всех 684 супермаркетов, за которые вы отвечаете.

в) Изменчивость количества нарушений в универмагах из генеральной совокупности.

г) Изменчивость количества нарушения в универмагах, измеренная вычисленным вами стандартным отклонением.

д) Стандартное отклонение вашего выборочного среднего.

е) Стандартная ошибка вашего выборочного среднего.

6. Постройте случайную выборку без возврата объемом 3 из следующей очень небольшой совокупности фирм: IBM, GM, Ford, Shell, HP, Boeing и ITT. Используйте следующую последовательность случайных цифр: 5887053671352339.

7. Постройте случайную выборку без возврата объемом 4 из следующей группы металлообрабатывающих корпораций: Gillette, Crown Cork & Seal, MASCO, Tyco Laboratories, Illinois Tool Works, McDermott, Ball, Stanley Works, Harsco, Hillenbrand Industries, Newell, Snap-on Tools, Danaher, Silgan, Robertson-Ceco и Barnes Group. Используйте таблицу случайных чисел, начните в строке 28 со столбца 7.

8. Из генеральной совокупности из полученных 681 счета извлеките случайную выборку из 3 номеров счетов. Используйте таблицу случайных чисел, начните в строке 6 со столбца 2.

9. Из генеральной совокупности 86 поставщиков постройте случайную выборку из 4 фирм. Начните в таблице случайных чисел со столбца 4 в строке 30.

10. Постройте случайную выборку контрактов объемом 5 из генеральной совокупности, включающей 362 контракта с перерасходом. Начните отсчет со строки 13 в столбце 5 таблицы случайных чисел.

11. Постройте случайную выборку объемом 8 из генеральной совокупности просроченных счетов за электроэнергию объемом 500. Начните со строки 17 в столбце 5 таблицы случайных чисел.

12. Для большой совокупности банковских счетов среднее баланса счетов составляет $500, а стандартное отклонение — $120. Определите стандартное отклонение среднего баланса счетов для группы из восьми счетов, отобранных независимо друг от друга.

13. Средняя производительность для генеральной совокупности составляет 35, стандартное отклонение для генеральной совокупности — 10, объем выборки равен 15. Определите стандартное отклонение общей суммы для случайной выборки.

14. Имеется восемь машин, работающих независимо друг от друга. Средняя производительность каждой машины — 20,3 тонны в день, стандартное отклонение составляет 1,4 тонны в день. Вычислите приближенное значение неопределенности в средней дневной производительности этих восьми машин. Используйте необходимую меру.

15. На вашей фабрике работает 40 одинаковых станков, имеющих среднюю производительностью 90 изделий в день со стандартным отклонением 35 изделий. Можно считать, что все станки работают независимо друг от друга. Рассмотрите случайную переменную "средняя дневная производительность одного станка на завтрашний день".

    а) Найдите среднее этой случайной переменной. Сравните это значение со средней производительностью отдельного станка.

    б) Определите стандартное отклонение этой случайной переменной. Сравните это значение со стандартным отклонением производительности отдельного станка.

    в) Какое распределение (приблизительно) имеет эта случайная переменная? Как вы это установили?

    г) Определите приблизительно вероятность того, что завтра средняя дневная производительность одного станка будет составлять от 95 до 100 изделий.

16. На упаковку завтрака с овсяной кашей быстрого приготовления нанесена следующая надпись: "Чистый вес — 20 унций, расфасовано в соответствии с весом, а не с объемом; во время перевозки возможна утруска". Однако фактически вес каждой упаковки не всегда точно равен 20 унциям — существуют случайные отклонения от этого веса. Опираясь на прошлые наблюдения, предположим, что средний вес составляет 20,04 унции, стандартное отклонение — 0,15 унции и распределение приближается к нормальному. Рассмотрим средний вес 30 упаковок, отобранных случайно и независимо.

    а) Каким будет среднее значение этой случайной величины?

    б) Какова изменчивость этой случайной величины?

    в) Какова приблизительно вероятность того, что средний вес будет меньше 20 унций?

17. Фермер владеет 5 одинаковыми кукурузными полями. Урожай каждого из этих полей имеет нормальное распределение и составляет в среднем 80 000 бушелей зерна со стандартным отклонением 15 000 бушелей. Определите вероятность того, что для пяти полей средний урожай с одного поля составит 88 000 бушелей.

18. Среднее генеральной совокупности составляет $65, стандартное отклонение генеральной совокупности — $30. Определите вероятность того, что среднее значение 35-ти случайно выбранных сделок будет находиться в диапазоне от $55 до $60. Можно принять, что распределение генеральной совокупности приближается к нормальному.

19. Результаты анализа возможного развития проекта в соответствии с четырьмя сценариями приведены в табл. 8.6.1. Предположим, вы имеете 40 таких проектов, которые финансируются независимо друг от друга. Определите (приближенно) вероятность того, что средняя прибыль одного проекта составит от $5 000 000 до $6 000 000.

20. Вы принимаете заказы по телефону в отделе торговли по каталогам. Средний размер одного заказа составляет $38,63 со стандартным отклонением $13,91. Можно считать, что заказы поступают независимо друг от друга.

    а) Можете ли вы определить, исходя только из этой информации вероятность того, что один телефонный звонок принесет вашей фирме заказ на сумму больше $40. Почему да или почему нет?

    б) Ожидают, что завтра оператор обработает 110 телефонных заказов. Определите среднее и стандартное отклонения результата дневной работы оператора.

    в) Каким будет приблизительное распределение вероятности для полного (суммарного) заказа в пункте "б"? Как вы это вычислите?

    г) Определите (приближенно) вероятность того, что завтра (см. п. "б") оператор примет заказов на общую сумму больше $3300.

    д) Определите (приблизительно) вероятность того, что завтра средняя стоимость заказа (см. п. "б") будет в пределах от $27 до $29.

21. Этим вечером ваш ресторан должен обслужить 50 заказов на ужин. Примем, что средний размер заказа составляет $60, стандартное отклонение равняется $40 и распределение немного смещено в сторону более дорогих заказов.

Таблица 8.6.1

Сценарий	Вероятность	Прибыль или убытки (млн дол.)
Плохой	0,10	–10
Более или менее удовлетворительный	0,15	2
Довольно хороший	0,50	5
Отличный	0,25	15

а) Определите среднее и стандартное отклонение стоимости всех 50 заказов.

б) Определите среднее и стандартное отклонение для среднего стоимости всех 50 заказов.

в) Какие дополнительные допущения необходимы, чтобы вы смогли сделать вывод о том, что сумма стоимости всех 50 заказов имеет приблизительно нормальное распределение?

г) Определите вероятность того, что сумма всех 50 заказов будет больше 3100 дол. в предположении, что имеет место нормальное распределение.

д) Определите вероятность того, что среднее всех 50 заказов будет лежать в пределах от $58 до $65, приняв, что имеет место нормальное распределение.

22. Средний размер заказа ваших потребителей составляет $2601 со стандартным отклонением $1275. Вы хотите знать, что произойдет, если завтра 45 ваших обычных клиентов одновременно и независимо друг от друга разместят заказы.

а) Определите среднее общей суммы завтрашних заказов.

б) Вычислите стандартное отклонение общей суммы завтрашних заказов.

в) Далее (для всех остальных пунктов данной задачи) считайте, что общая сумма завтрашних заказов имеет нормальное распределение. Почему такое допущение обоснованно, даже если распределение заказов отдельных клиентов несколько асимметрично?

г) Определите вероятность того, что общая сумма заказов будет равна или превысит сумму $105 000, которая для вас является точкой самоокупаемости.

д) Определите вероятность того, что общая сумма заказов превысит $135 000 (это будет по-настоящему удивительный день!).

е) Определите вероятность того, что общая сумма заказов составит (как в обычные дни) от $110 000 до $125 000 дол.

ж) Определите вероятность необычного дня, когда общая сумма заказов будет либо меньше $100 000 либо больше $135 000.

з) Какова вероятность, что завтра средняя стоимость одного заказа составит от $2450 до $2750?

23. У вас есть фабрика с 40 одинаковыми станками. Средняя дневная производительность каждого станка составляет 100 изделий со стандартным отклонением 15 изделий. Можно считать, что все станки работают независимо друг от друга. Рассмотрите среднюю завтрашнюю дневную производительность на один станок, которая является случайной величиной.

а) Найдите среднее этой случайной величины. Сравните его со средним для отдельного станка.

б) Определите стандартное отклонение этой случайной величины. Сравните его со стандартным отклонением для отдельного станка.

в) Какое приблизительно распределение вероятности имеет эта случайная величина? Почему вы так считаете?

г) Определите (приблизительно) вероятность того, что завтра средняя дневная производительность на один станок составит более 102 изделий.

д) Определите (приблизительно) вероятность того, что завтра средняя дневная производительность на один станок составит от 97 до 103 изделий.

24. Рассмотрим прибыль как процент от дохода компаний, выпускающих промышленное и сельскохозяйственное оборудование (табл. 8.6.2).

а) Постройте основу выборки, рассматривая данный список как генеральную совокупность крупных фирм, выпускающих промышленное и сельскохозяйственное оборудование.

б) Постройте выборку фирм объемом 10, начав в строке 13 со столбца 2 таблицы случайных чисел.

в) Вычислите выборочное среднее.

Таблица 8.6.2. Прибыль компаний, выпускающих промышленное и сельскохозяйственное оборудование

Фирма	Прибыль как процент от дохода, %	Фирма	Прибыль как процент от дохода, %
Caterpillar	8,8	Detroit Diesel	1,4
Deere	7,5	Aeroquip-Vickers	4,8
Dresser Industries	4,3	Crane	5,5
Ingersoll-Rand	5,4	Cincinnati Milacron	4,2
Case	6,7	Cooper Cameron	7,8
American Standard	1,6	Tecumseh Products	5,8
Cummins Engine	3,8	Smith International	6,5
Black & Decker	4,6	Stewart & Stevenson Services	2,1
Dover	8,9	Unova	–12,0
Parker Hannifin	6,7	United States Filter	3,4
Baker Hughes	2,6	Briggs & Stratton	4,7
AGCO	5,2	Nortek	1,7
York International	1,5	Lincoln Electric	7,4
Harnischfeger Industries	4,5	Kennametal	6,2
Western Atlas	–2,2	Teleflex	6,1
Timken	6,5	Pall	6,3
Premark International	4,3	Toro	3,3
Nacco Industries	2,8		

Данные по состоянию на 1.07.1999 г., взяты по адресу
`http://www.pathfinder.com/fortune/fortune500/ind13.html`

г) Вычислите стандартную ошибку среднего как с поправкой на конечность генеральной совокупности, так и без нее.

д) Напишите один абзац текста с интерпретацией стандартной ошибки.

е) Вычислите среднее генеральной совокупности. (*Замечание.* В реальной жизни обычно это сделать невозможно!)

ж) В одном абзаце кратко объясните взаимосвязь между выборочным средним, средним генеральной совокупности и стандартной ошибкой.

25. Рассмотрим темпы роста доходов на одну акцию за 10-летний период для пяти наиболее крупных фирм, производящих мебель (см. табл. 8.6.3). Считайте данный список генеральной совокупностью (очень малой!), включающей всего лишь $N = 5$ элементов. Рассмотрим выборку объемом $n = 2$.

а) Составьте список всех выборок объемом 2, которые могут быть извлечены. (*Подсказка.* Существует 10 *таких выборок.*) Для каждой выборки вычислите среднее.

б) Постройте гистограмму десяти выборочных средних из п. "а". Это будет *выборочное* распределение выборочного среднего.

в) Постройте случайную выборку объемом 2 из генеральной совокупности, начав в строке 26 со столбца 4 таблицы случайных чисел. Вычислите среднее этой выборки.

г) Покажите, где на гистограмме выборочного распределения (п. "б") находится среднее, полученное в п. "в".

д) Напишите один абзац текста, поясняющий, почему "построение случайной выборки из генеральной совокупности и вычисление среднего для этой выборки" приводит, по сути, к тому же результату, что и "извлечение случайного значения из выборочного распределения выборочных средних".

26. Экономисты часто делают прогнозы будущих событий. *Рассмотрим* сделанный во время опроса экономистов в июле 1998 г. прогноз на 31 декабря 1998 г. ставки трехмесячных векселей казначейства США (табл. 8.6.4).

а) Вычислите *среднее и стандартное отклонение.* Кратко поясните полученные результаты.

б) Вычислите стандартную ошибку среднего. Поясните полученное значение, рассматривая данный список как случайную выборку из более обширного списка экономистов.

Таблица 8.6.3. Темпы роста доходов на одну акцию за 10-летний период

Фирма	Рост доходов на одну акцию, %
Johnson Controls	2
Avery Dennison	3
Leggett & Platt	8
Herman Miller	–8
Hon Industries	9

Данные взяты из "Fortune 500", *Fortune*, April 20, 1992, p. 274.

в) Через шесть месяцев после того, как был сделан прогноз, процентная ставка для трехмесячных векселей казначейства фактически составила 4,46%. Сравните средний прогноз с этим фактическим значением.

г) На сколько величин стандартной ошибки отличается выборочное среднее от фактического результата (4,46%)? Удивлены ли вы таким большим отличием?

Таблица 8.6.4. Экономические прогнозы относительно процентных ставок трехмесячных векселей казначейства

Экономист	Прогноз процентной ставки на 31.12.98, сделанный в июле 1998 г., %	Экономист	Прогноз процентной ставки на 31.12.98, сделанный в июле 1998 г., %
Allyn	4,80	Lonski	5,25
Angell	5,10	McCulley	4,66
Berner	4,90	McDevitt	5,10
Berson	5,10	Moskowitz	5,20
Blitzer	4,75	Mueller	5,50
Braverman	5,10	Perna	4,75
Brown	5,00	Platt	5,34
Brusca	5,45	Ramirez	5,10
Bussmann	5,00	Ralajczak	5,17
Cahn	5,15	Resler	5,00
Coons	5,20	Reynolds	5,30
Cosgrove	5,20	Rippe	5,30
Daane	5,50	Shilling	4,75
Dubley	5,20	Sinai	5,08
Englund	5,40	Smith	4,25
Foster	5,20	Sohn	5,25
Gallagher	5,00	Steinberg	4,90
Harris	5,20	Sterne	4,50
Herrick	5,50	Swonk	5,10
Hoffman	5,05	Synott	5,10
Hummer	5,41	Walter	4,78
Hyman	5,00	Wesbury	4,65
Hymans	5,14	Williams	5,45
Karl	5,30	Worseck	4,80
Laufenberg	5,20	Wyss	5,10
Levy	5,00	Yardeni	5,00
Littmann	5,00	Zandi	5,10

Данные взяты из Mitchell Ford, C., "U. S. Economy is Seen at Slower Pace in 1999," *The Wall Street Journal*, January 4, 1999, p. A2.

д) Объясните, почему ошибка прогноза (среднее прогноза минус фактическое значение) не должна быть приблизительно равна стандартной ошибке. Для этого определите среднее генеральной совокупности и покажите, что это среднее *не* то же самое, что фактическое значение.

27. Имеется список сумм (в долларах) последних счетов:

$994, $307, $533, $443, $646, $148, $307, $524, $71, $973, $710, $342, $494

а) Вычислите среднее. Что означает это число?

б) Вычислите стандартное отклонение. Что означает это число?

в) Вычислите стандартную ошибку. Что означает это число?

г) Ожидается, что в следующем месяце будут разосланы еще 500 подобных счетов. Сделайте прогноз относительно общей суммы, на которую будут выписаны эти дополнительные счета.

28. Имеется выборка из 200 пенсионеров поселка. Среднее значение возраста для этой выборки пенсионеров составляет 69,8 лет, а стандартное отклонение — 9,2 года. Ваш друг утверждает, что выборочное среднее отличается от среднего генеральной совокупности на 9,2. Прав ли он? Почему вы так считаете?

29. Определите стандартную ошибку среднего для следующего набора данных, представляющих собой уровень качества сельскохозяйственной продукции:

16,7; 17,9; 23,5; 13,8; 15,9; 15,2; 12,9; 15,7

30. Случайная выборка из 50 карточек пациентов, недавно посетивших клинику, показывает, что в среднем один пациент платит за визит к врачу $53,01 и стандартное отклонение составляет $16,48.

а) Определите стандартную ошибку среднего и поясните ее значение.

б) Вы считаете, что для обоснованного финансового планирования эта стандартная ошибка слишком велика. Определите ожидаемую стандартную ошибку для выборки из 200 человек при условии, что стандартное отклонение не изменится.

31. Определите среднее и стандартную ошибку суммы денег, потраченных вашими постоянными клиентами на покупку ваших товаров в прошлом месяце, рассматривая данные из задачи 2 главы 4, в качестве выборки заказов клиентов.

32. Определите среднее и стандартную ошибку для прочности хлопчатобумажной пряжи, используемой на ткацкой фабрике, исходя из данных задачи 15 главы 4.

33. Определите среднее и стандартную ошибку веса конфеты до вмешательства в процесс производства, используя данных задачи 11 главы 5.

34. Опрос 823 случайно отобранных взрослых жителей США показал, что 63% из них поддерживают текущую политику правительства. Найдите показатель, который приблизительно охарактеризовал бы отличие этого выборочного процента от значения, которое могло бы быть получено при опросе всего взрослого населения США.

35. Проводится исследование, посвященное опознанию потребителями вашей торговой марки. Из 763 случайно отобранных для опроса человек 152 не смогли опознать вашу продукцию.

а) Оцените процентное содержание людей в генеральной совокупности (из которой была взята эта выборка), которые не смогут опознать вашу продукцию.

б) Определите стандартную ошибку оценки, найденной в п. "а", и кратко поясните полученный результат.

36. На основе тщательного изучения выборки объемом 868, извлеченной из 11013 хранящихся на складе контейнеров, вы обнаружили, что 3,6% контейнеров не готовы к отгрузке.

а) Определите стандартную ошибку, связанную с приведенной выше оценкой процента, и поясните полученный результат.

б) Были бы вы удивлены, узнав, что в действительности 4% из 11013 контейнеров не готовы к отгрузке? Почему?

в) Были бы вы удивлены, узнав, что в действительности 10% из 11013 контейнеров не готовы к отгрузке? Почему?

37. Из списка 729 участников круиза случайным образом было опрошено 25 человек. Из них 21 человек заявил, что остались "очень довольны" предоставленным обслуживанием.

а) Какой процент выборки дал ответ "очень довольны"?

б) Если бы была возможность опросить все 729 человек, то насколько отличался бы процент довольных обслуживанием от полученного в п. "а"? Чтобы ответить на этот вопрос, пожалуйста, определите, какую статистику необходимо использовать, и вычислите ее значение.

38. Газета провела опрос 1487 своих читателей и оказалось, что 42,3% из них готовы отдать свои голоса за данного кандидата. Выборы состоятся через три недели.

а) Какой приблизительно процент всех жителей ответил бы о готовности отдать голос за данного кандидата, если бы все население было опрошено при тех же условиях?

б) Приведите две причины, почему фактический результат выборов может отличаться от этих 42,3% больше, чем на величину стандартной ошибки?

39. Счета фирмы были сгруппированы следующим образом: 56 крупных, 956 средних и 16 246 мелких счетов. Каждый счет имеет *балансовую стоимость* (которую вам предоставили), представляющую количество денег, которое, как *полагают*, находится на данном счете. Каждый счет также имеет *контрольную стоимость* (получение которой требует времени и усилий), показывающую количество денег, находящихся на данном счете *фактически*. Вы работаете с аудиторами, готовящими финансовые отчеты. Решено проверить 56 крупных, 15% средних и 2% мелких счетов. Совокупная ошибка (разность между балансовой стоимостью и контрольной) составила для крупных счетов $15 018, для средних — $1 165 и для мелких — $792.

Стандартные отклонения ошибок составили соответственно: $968,62; $7,12 и $5,14. (*Подсказка*. Не путать *ошибку*, которую определяют для каждого счета, со *стандартной ошибкой* среднего.)

а) Определите выборочное среднее ошибки, приходящейся на один счет в каждой из трех типичных групп (страт) счетов.

б) Объедините эти три средние ошибки, чтобы найти стратифицированную выборочную среднюю оценку средней ошибки, приходящейся на один счет, для генеральной совокупности.

в) Определите стандартную ошибку своей оценки, полученной в п. "б", "в" учетом и без учета поправки на конечность генеральной совокупности.

г) Объясните (уточненное) значение стандартной ошибки в терминах (неизвестного) значения средней ошибки, приходящейся на один счет, для генеральной совокупности.

40. В рамках исследования, связанного с розничной торговлей обувью, среди случайно выбранных потребителей из четырех городов провели опрос. Каждый потребитель сообщил имеющееся у него количество пар обуви (каждая строка представляет один город) (табл. 8.6.5).

а) Оцените среднее количество пар обуви для общего населения всех четырех городов.

б) Определите стандартную ошибку для оценки в п. "а".

в) Вычислите стандартную ошибку с учетом и без учета поправки на конечность генеральной совокупности и сравните их. Почему эти два полученных значения столь близки?

Таблица 8.6.5

Размер генеральной совокупности (население)	Размер выборки	Среднее выборки (количество пар обуви)	Стандартное отклонение выборки
3 638 815	200	13,77	13,57
6 899 665	200	12,72	12,11
9 608 853	250	8,79	12,34
709 212	200	10,43	14,99

Упражнения с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении А. Рассматривайте эту базу данных как интересующую вас генеральную совокупность.

1. Покажите, что эта база данных создана как основа выборки. В частности, покажите, как ее можно использовать, чтобы получить доступ к информации об отдельном служащем.
2. Используя таблицу случайных чисел, постройте случайную выборку объемом 10 без возврата. В качестве начальной точки в таблице возьмите строку 23, столбец 7.

а) Перепишите номера служащих, которые попали в вашу выборку.

б) Вычислите среднюю заработную плату для вашей выборки и интерпретируйте полученное число.

в) Вычислите стандартное отклонение заработной платы в данной выборке и интерпретируйте полученное число.

г) Вычислите стандартную ошибку заработной платы для вашей выборки и интерпретируйте полученное число. В частности, почему стандартная ошибка отличается от стандартного отклонения, найденного в предыдущем пункте этого упражнения?

3. Продолжайте работу с выборкой из предыдущего упражнения.

а) Определите среднее значение заработной платы для генеральной совокупности. (*Замечание*. В реальной жизни обычно нет возможности определить среднее для генеральной совокупности. В данном случае мы как бы имеем возможность подглядывать "из-за кулис".)

б) Сравните среднее значение заработной платы для генеральной совокупности с выборочным средним. В частности, на сколько величин стандартной ошибки отличаются эти два значения?

в) Вычислите стандартное отклонение заработной платы для генеральной совокупности и интерпретируйте полученное число.

г) Сравните стандартное отклонение заработной платы для генеральной совокупности с выборочным стандартным отклонением.

д) Для выборочного среднего значения заработной платы найдите стандартное отклонение генеральной совокупности и интерпретируйте полученное число. Сравните его со стандартной ошибкой, полученной из данных выборки.

е) Запишите полученные числа в форме таблицы со столбцами "генеральная совокупность" и "выборка" и строками "выборочное среднее и среднее генеральной совокупности", "стандартное отклонение для отдельных служащих", а также "стандартное отклонение и стандартная ошибка выборочных средних для 10 служащих".

4. Выполните упражнение 2, рассматривая вместо заработной платы возраст служащих.
5. Выполните упражнение 2, рассматривая вместо заработной платы стаж работы служащих.
6. Выполните упражнение 3, рассматривая вместо заработной платы возраст служащих.
7. Выполните упражнение 3, рассматривая вместо заработной платы стаж работы служащих.
8. Продолжайте работу с выборкой из упражнения 2.

а) Определите биномиальное X для переменной, определяющей пол служащего (подсчитав число служащих женского пола), и интерпретируйте полученное число.

б) Определите стандартную ошибку X и интерпретируйте ее.

в) Для биномиального X вычислите среднее генеральной совокупности.

г) Насколько наблюдаемое в вашей выборке значение X отличается от соответствующего среднего в генеральной совокупности?

д) Насколько велика эта разница по сравнению со стандартной ошибкой X?

9. Выполните упражнение 8, подставив биномиальную долю p вместо X.

Проекты



1. В настоящее время в Internet часто можно получить информацию о финансовом положении отдельных фирм — либо в виде отчетов (например, на электронной странице журнала *Fortune*), либо на электронных страницах фирм (часто под заголовком "Для инвесторов"). Рассмотрите какой-то важный показатель, например "прибыль как процент от дохода", который содержательно можно сравнивать для крупных и небольших фирм.

   а) Определите совокупность интересующих вас фирм и создайте основу выборки.

   б) Сделайте случайную выборку из 10 фирм. Найдите данные для этих фирм.

   в) Вычислите среднее и стандартную ошибку.

   г) Определите (приблизительно), насколько отличается ваше среднее от среднего значения, рассчитанного для всех фирм из основы выборки.

   д) Напишите абзац текста, подытоживающий все, что вы узнали о выборочных статистиках и о фирмах в вашей генеральной совокупности.

2. Ваша фирма планирует маркетинговую стратегию для "нового и улучшенного" потребительского товара. Ваше рекламное агентство подготовило 5 вариантов телевизионной рекламы, и вы должны выбрать 2 из них. Просмотрев все материалы, вы поняли что некоторые рекламные ролики больше обращены к женщинам, а не к мужчинам. Прежде чем вложить $1 800 000 в рекламную кампанию, ваш руководитель хотел бы получить больше информации о реакции потребителей на эти рекламные ролики. Напишите своему руководителю служебную записку с предложением о том, как собрать необходимую для принятия этого решения информацию. Обязательно охватите следующие темы: случайная выборка, стратифицированная случайная выборка, пробное исследование.

3. Определите ситуацию, относящуюся к ваше работе или экономической деятельности, в которой могла бы оказаться полезной статистическая выборка.

   а) Опишите генеральную совокупность и определите способ извлечения выборки.

   б) Определите интересующий вас параметр генеральной совокупности и укажите, как выборочная статистика может помочь в определении неизвестного параметра.

   в) Объясните понятие выборочного распределения этой статистики для вашего конкретного примера.

Ситуация для анализа

Можно ли извлечь пользу из этого исследования?

"Меня беспокоит мысль о том, что нельзя использовать новую случайную выборку только потому, что кому-то не нравятся результаты первого исследования. Пожалуйста, расскажите мне подробнее о проделанной работе". Ваш голос тверд и уверен, вы пытаетесь понять, что произошло в действительности, и надеетесь извлечь полезную информацию без дополнительных расходов на новое исследование.

"Не то, чтобы нам не понравились *результаты* первого исследования, — ответил Р. Л. Стигманс, — но было опрошено только 54% респондентов. Мы даже не посмотрели на планируемые ими расходы, когда принималось решение (снова взять выборку). Поскольку мы просто планировали получить ответы от почти всех 400 первоначально отобранных лиц, мы затем случайно выбрали еще 200 человек и также опросили их. Это и была вторая выборка". В этом месте, размышляя, чтобы еще добавить к этому, вы вздыхаете.

"Затем у Э. С. Элдриджа появилась идея о том, как продолжить работу с теми, кто не ответил. Мы послали им другой полный вопросник вместе с хрустящей долларовой банкнотой и письмом, в котором объяснили, как важны для нас их ответы, чтобы спланировать производство. Сработало достаточно хорошо. Тогда, конечно, мы сделали такое же напоминание и для второй выборки".

"Если я правильно вас понял, — отвечаете вы, — у вас есть две выборки: одна из 400 человек, вторая из 200. Для каждой из них есть ответы, полученные без напоминания и после напоминания. Я правильно вас понял?"

"Да, конечно, но есть еще пробное исследование 12 человек, опрошенных в офисах этажом ниже и на другой стороне улицы. Нам хотелось бы включить и эти данные, усреднить их с остальными, потому что это была сложная работа в начале исследования и теперь жаль просто выбросить эти результаты. Что мы действительно хотим узнать — это средний размер расходов с точностью до сотни долларов".

С этого момента вы понимаете, что у вас уже достаточно информации, чтобы оценить ситуацию и рекомендовать либо выполнять необходимую оценку, либо проводить новое исследование. Ниже приведены дополнительные подробности опроса 8391 респондента для определения планируемых в следующем квартале расходов.

	Пробное исследование	Первая выборка	Вторая выборка	Обе выборки	Объединенные результаты
Первоначальная почтовая рассылка					
Отослано	12	400	200	600	612
Ответили	12	216	120	336	348
Среднее	$39 274, 89	$3 949,40	$3 795,55	$3 894,45	$5 114,47
Стандартное отклонение	$9 061,91	$849,26	$868,39	$858,02	$6 716,42

Окончание таблицы

	Пробное исследование	Первая выборка	Вторая выборка	Обе выборки	Объединенные результаты
Почтовая рассылка напоминания					
Отослано	0	184	80	264	264
Ответили	0	64	18	82	82
Среднее		$1 238,34	$1 262,34	$1 243,60	$1 243,60
Стандартное отклонение		$153,19	$156,59	$153,29	$153,29
Первоначальная рассылка и напоминание вместе					
Отослано	12	400	200	600	612
Ответили	12	280	138	418	430
Среднее	$39 274,89	$3 329,73	$3 465,13	$3 374,43	$4 376,30
Стандартное отклонение	$9 061,91	$1 364,45	$1 179,50	$1 306,42	$6 229,77

Вопросы для обсуждения

1. Согласны ли вы с тем, что идея использовать вторую выборку была хорошей?
2. Согласны ли вы с тем, что идея рассылать напоминания по почте была хорошей?
3. Как вы можете объяснить полученные в результате различия в значениях средних?
4. Можно ли извлечь пользу из приведенных здесь результатов? Какие из них полезны? Достаточно ли этих данных или необходимо проводить дальнейшие исследования?

ГЛАВА 9

Доверительные интервалы: допущение о неточности оценок

Существуют два пути получения более высокой прибыли: увеличение дохода и уменьшение расходов. В медицинском страховании управлять уровнем доходов сложно, так как страховые компании и правительство устанавливают максимальную сумму, которую они могут заплатить в качестве компенсации при данном диагнозе. Рассмотрим больницу, менеджеры которой пытаются найти разумный ответ на вопрос: "Как много денег в расчете на одного пациента кардиохирургии мы можем заработать или потерять в долгосрочном плане?". Тщательный анализ выборки из медицинских и финансовых отчетов о 35 пациентах кардиохирургии показал, что средний доход составляет $390,26 со стандартным отклонением $450,56. Итак, об этих 35 пациентах информации достаточно. Но что известно о пациентах кардиохирургии в целом? Вас интересуют не только эти пациенты, а что произойдет в будущем. Вспомним, что стандартная ошибка $450,56/\sqrt{35} = 76,16$ приближенно показывает, насколько отличается среднее выборки, равное $390,26, от среднего генеральной совокупности, из которой эта выборка взята. Однако необходимо двигаться дальше, поскольку такой "аппроксимации" недостаточно. Вам необходимо сформулировать определенное, заслуживающее доверие, точное утверждение. Именно для этой цели используют доверительные интервалы. В данном случае доверительный интервал (способы его вычисления будут описаны ниже) позволяет утверждать следующее.



Мы на 95% уверены, что средний доход в расчете на одного пациента для совокупности, из которой взята выборка из 35 пациентов, находится в пределах от $235,51 до $545,01.

Это утверждение о генеральной совокупности определяет область вокруг выборочного значения среднего, $390,26, является действительно точным утверждением и отражает случайность выборки. Мы не изучали всю генеральную совокупность (которая включает намного большее количество пациентов кардиохирургии), и у нас нет в этом необходимости. Тем не менее мы можем сделать такое же утверждение, как могли бы получить, если бы потратили деньги на анализ и обобщение данных из старых отчетов. Доверительный интервал демонстрирует важную связь между доступным для нас исследованием 35 пациентов и гораздо большей генеральной совокупностью, и эта связь выходит за пределы приближенной интерпретации стандартной ошибки.[1]

Практическим следствием такого доверительного интервала является то, что при распространении выводов за пределы данных об исследованных пациентах значение выборочного среднего $390,26 оказывается не настолько точным, как казалось вначале. Может быть, ошибка в размере $100 на одного пациента (причем отклонения могут быть в обоих направлениях). Почему ошибка так велика? Причина в изменчивости (одни пациенты приносят прибыль больше, чем другие) и в малом объеме выборки (исследовав данные более чем о 35 пациентах, мы получим более точную оценку среднего генеральной совокупности).

Доверительный интервал можно также использовать, чтобы показать, насколько точно выраженная в процентах доля признака в выборке отражает интересующую нас долю признака в совокупности. Например, результаты маркетингового опроса 150 человек, случайно отобранных из вашей целевой группы, показали, что 46 человек, или 30,7%, знают вашу торговую марку. Вы, конечно же, не верите, что точно 30,7% всей целевой группы знают вашу торговую марку, поскольку вам известно, что случайность процесса построения выборки приводит к ошибке, которая приблизительно равна одной стандартной ошибке. В данном случае стандартная ошибка $S_p = 3,76$ процентных единиц показывает приблизительное различие между выраженными в процентах долями в выборке и в генеральной совокупности. Доверительный интервал, который вычисляют с помощью приведенных далее в этой главе методов, формализует это понятие *приблизительной разности* и позволяет сделать такое заключение.

Мы уверены на 95%, что доля людей, которым известна наша торговая марка, в нашей целевой группе (в генеральной совокупности) находится где-то между 23,3 и 38,0%.

Цель использования доверительных интервалов заключается в том, чтобы по возможности избавиться от неопределенности и сделать как можно более точный вывод. Вероятность дает нам возможность формулировать точные утверждения в условиях неопределенности. Статистика дает возможность извлекать необходимую информацию из данных выборки. Процесс обобщения данных выборки, который приводит к вероятностным утверждениям о всей генеральной совокупности, называют **статистическим выводом**. В частности, **доверительным интервалом** называют такой вычисленный на данных интервал, который с *известной вероятностью* содержит интересующий нас неизвестный параметр генеральной совокупности, и эта вероятность определяется с учетом случайного эксперимента, который начинается с извлечения случайной выборки. Таким образом, определить доверительный интервал — это лучшее, что можно сделать в условиях неопределенности: это точное вероятностное утверждение вместо неясных замечаний типа: "Мы не уверены, но..." или "Это значение, вероятно, близко к...".

[1] А что произойдет в том случае, если все записи о пациентах уже занесены в компьютер? Тогда легче будет анализировать более крупные выборки. Однако всегда будет расхождение между пациентами, которых вы наблюдали (выборка), и тем, что будет, вероятно, происходить в будущем. Доверительный интервал будет оставаться полезным и в таком случае, поскольку он фиксирует случайный компонент этой разности. Однако следует также учитывать и систематические изменения на рынке или в технологиях.

Доверительные интервалы используют часто, и ниже приведен краткий предварительный обзор их полезных свойств. Есть возможность выбирать вероятность утверждения. Эту вероятность называют **доверительным уровнем** (используют также термины "коэффициент доверия" и "доверительная вероятность". — *Прим. ред.*). Традиционно его устанавливают равным 95%, но часто используют также значения 90, 99 и даже 99,9%. Платой за более высокий доверительный уровень является более широкий, а значит, и менее полезный интервал. Доверительный интервал для процентного содержания в генеральной совокупности можно легко вычислить, используя стандартную ошибку для биномиального распределения. В зависимости от необходимости можно использовать двусторонний (между двумя значениями) или односторонний (по крайней мере больше, чем некоторое значение) доверительный интервал. Как всегда, следует быть осторожным с не всегда декларируемыми явно, но необходимыми предварительными техническими условиями (в данном случае это нормальность и случайность выборки), поскольку если эти условия не удовлетворяются, то сформулированные на основе доверительных интервалов выводы будут неверными. Необходимо также тщательно различать вероятность 95% для *процесса* построения доверительного интервала и 95% доверительный уровень для конкретного вычисленного интервала.

Сформулируем приблизительное универсальное утверждение о доверительном интервале, которое применяют во многих ситуациях. Если вы с помощью соответствующей несмещенной оценки оценили параметр генеральной совокупности и вычислили соответствующую стандартную ошибку этой оценки, то утверждение о доверительном интервале (в обобщенном виде) можно сформулировать следующим образом.

Приблизительное утверждение о доверительном интервале

Мы на 95% уверены, что параметр генеральной совокупности находится между значением оценки *минус* две стандартные ошибки и значением оценки *плюс* две стандартные ошибки.

Следует помнить, что значение нормально распределенной переменной находится в пределах двух стандартных отклонений от своего среднего приблизительно в 95% случаев; вот откуда (отчасти косвенно) возникли эти значения в обобщенной формулировке утверждения о доверительном интервале.

Насколько широко можно применять понятие доверительного интервала? По существу, любое число, которое вы встречаете в газетах, каких-то ваших конфиденциальных стратегических внутренних документах или в телепередачах, является оценкой некоторого важного значения. По сути, все эти оценки имеют соб-

ственные "личные" стандартные ошибки, характеризующие их точность. Знание этих двух величин (оценка и ее стандартная ошибка) позволяет использовать указанное выше приблизительное утверждение о доверительном интервале. Однако далее мы рассмотрим ряд деталей, особенностей и ограничений, которые нужно учитывать в конкретных случаях.

9.1. Доверительный интервал для среднего значения и для доли признака в генеральной совокупности

Мы только что извлекли выборку и с целью оценить среднее генеральной совокупности вычислили выборочное среднее 1. Пусть нам известно (обычно неизвестное) среднее значение генеральной совокупности, μ, как это показано на рис. 9.1.1.

Расстояние между средним выборки и средним совокупности (ошибка оценивания) *не зависит* от того, будем ли мы использовать в качестве начальной точки измерения этого расстояния выборочное среднее или среднее генеральной совокупности. Иными словами, проведя измерение в единицах стандартной ошибки от среднего совокупности, мы получим тот же результат, который нам даст измерение в единицах стандартной ошибки от выборочного среднего. Это не тривиальная мысль. Поскольку среднее выборки известно, можно проводить измерение в единицах *известного* значения (стандартной ошибки), взяв в качестве начальной точки измерения другое *известное* значение (среднее выборки), и при этом получить

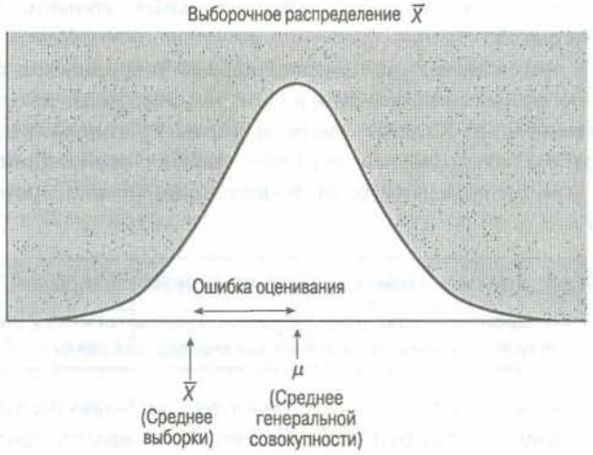


Рис. 9.1.1 Выборочное распределение $\overline{X}$ имеет своим центром значение μ. Расстояние между средним выборки и средним совокупности (ошибка оценивания) одно и то же, независимо от точки отсчета расстояния — от среднего выборки или от среднего совокупности. Доверительный интервал строят, взяв за точку отсчета известное значение $\overline{X}$, а не неизвестное значение μ

ту же взаимосвязь, как если бы для измерения в качестве начальной точки использовалось (неизвестное) среднее значение генеральной совокупности.

Интуитивные основания построения доверительного интервала следующие. Вспомним, что для нормально распределенной случайной величины вероятность оказаться в пределах двух стандартных отклонений от среднего составляет приблизительно 0,95.[2] Отсюда получаем следующее вероятностное утверждение.

Вероятность того, что выборочное среднее находится в пределах 1,960 "стандартных отклонений выборочного среднего" от среднего генеральной совокупности, равна 0,95.

Однако это утверждение основано на измерениях относительно *неизвестного* среднего генеральной совокупности, μ. Чтобы избежать этой проблемы, можно провести такие же измерения, но в качестве точки отсчета взять выборочное среднее. В таком случае получим следующее эквивалентное предыдущему вероятностное утверждение.

Вероятность того, что среднее генеральной совокупности находится в пределах 1,960 "стандартных отклонений выборочного среднего" от выборочного среднего, равна 0,95.

Это утверждение все еще включает неизвестный параметр генеральной совокупности, поскольку стандартное отклонение выборочного среднего равно $\sigma_{\overline{x}} = \sigma / \sqrt{n}$. Те, кто занимаются статистикой, часто вместо неизвестных параметров подставляют известные оценки этих параметров. Наше утверждение будет оставаться приблизительно справедливым, если вместо стандартной ошибки использовать значение $S_{\overline{x}} = S / \sqrt{n}$, представляющее наилучшую известную нам информацию о стандартном отклонении среднего выборки. Это приведет к следующему приблизительному вероятностному утверждению:

Вероятность того, что среднее генеральной совокупности находится в пределах 1,960 *стандартных ошибок* от выборочного среднего, *приблизительно* равна 0,95.

К сожалению, это только приблизительное вероятностное утверждение. Чтобы сделать его точным, воспользуемся таблицей t-распределения, открытого Стьюдентом (Student) и опубликованного в 1908 г.[3] Взяв из t-таблицы вместо 1,960 соответствующее критическое значение, получим вместо приблизительного следующее *точное* вероятностное утверждение:

Вероятность того, что среднее генеральной совокупности находится в пределах (критическое значение из t-таблицы) стандартных ошибок от выборочного среднего, равна 0,95.

Ценой, которую мы платим за замену неизвестного параметра генеральной совокупности ($\sigma_{\overline{x}}$) соответствующей выборочной оценкой (стандартной ошибкой $S_{\overline{x}}$, вычисленной на имеющихся данных), является то, что критическое значение из t-таблицы будет больше, чем 1,960, что, в свою очередь, приведет к более широкому, а значит, менее точному интервалу. Если n больше 40, при расчетах вручную можно в качестве приближенного критического значения взять 1,960,

[2] Можно проверить по таблице нормального распределения, что вероятность находится в пределах 1,960 стандартных отклонений и *точно* равна 0,95.

[3] "Стьюдент" — псевдоним У. С. Госсета (W. S. Gossett), управляющего пивоваренной компании Guinness Brewery. Он разработал этот важный метод для контроля за процессом и для улучшения технологии пивоварения.

хотя компьютерные программы, как правило, используют точное, несколько более высокое критическое значение из t-таблицы.

Чтобы получить практически используемый доверительный интервал из этого вероятностного утверждения, следует заменить слова "вероятность 0,95" на слова "доверительный уровень 95%". Это необходимо сделать в связи с тем, что доверительный интервал на практике формулируют в терминах значений, а не в терминах случайных переменных (более подробно об этом речь пойдет в разделе 9.3). В окончательном виде доверительный интервал (рис. 9.1.2) имеет следующий вид.

Точная формулировка доверительного интервала для среднего генеральной совокупности

Мы на 95% уверены, что среднее генеральной совокупности лежит между значением оценки минус t стандартных ошибок и значением оценки плюс t стандартных ошибок. Таким образом, мы на 95% уверены, что среднее совокупности μ находится где-то между

$$\overline{X} - t\ S_{\overline{X}} \text{ и } \overline{X} + t\ S_{\overline{X}},$$

где t взято из t-таблицы для двустороннего доверительного уровня 95%. Шансы на то, что среднее генеральной совокупности находится за пределами этого интервала, равны 5%.

Такой же подход принят при оценивании неизвестной доли (процента) свойства в генеральной совокупности, исходя из доли (процента) свойства в выборке. Это возможно потому, что доля p свойства в выборке представляет собой среднее $\overline{X}$ при условии, что случайная переменная X принимает значения 1 или 0, в зависимости от наличия или отсутствия изучаемого свойства (и, следовательно, процент в генеральной совокупности π также равен среднему совокупности μ). Например, результатом небольшого опроса группы из 5 человек, которым был задан вопрос: "Нравится ли вам цвет этого изделия?", может быть ряд значений 0, 1, 0, 0, 1, свидетельствующий о том, что цвет изделия нравится второму и пятому из опрошенных. Отсюда, $p = \overline{X} = 0{,}4$, или 40%. В соответствии с центральной предельной теоремой (которая позволяет использовать нормальное распределение в качестве аппроксимации для биномиального), если n велико, а π не слишком близко к 0 или 1, значения p будут приблизительно нормально распределены и доверительный интервал будет корректным.

Рассмотрим ситуацию, связанную с биномиальным распределением, когда оценивается неизвестная вероятность π некоторого события, на основе наблюдаемых X наступлений этого событий в результате n попыток (при большом

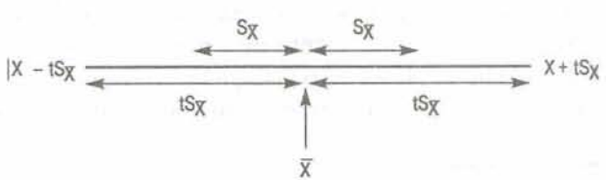


Рис. 9.1.2. 95% доверительный интервал находится в пределах t (приблизительно двух) стандартных ошибок, $S_{\overline{X}}$, в обе стороны от среднего выборки $\overline{X}$

значении n). Вспомним из главы 7, что для оценки π мы используем долю в выборке, равную $p = X/n$, и что стандартная ошибка p вычисляется по формуле $S_p = \sqrt{p(1-p)/n}$. В соответствии с формулировкой утверждения о доверительном интервале, используя в качестве параметра генеральной совокупности π (вместо μ), в качестве оценки p (вместо $\overline{X}$) и в качестве стандартной ошибки S_p (вместо $S_{\overline{X}}$), получим следующий доверительный интервал (рис. 9.1.3).[4]

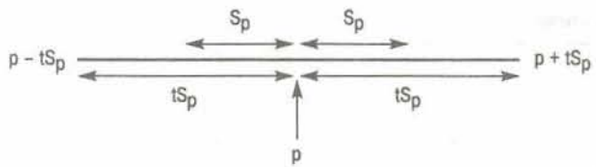


Рис. 9.1.3. Для биномиальной доли π 95% доверительный интервал находится в пределах t (приблизительно двух) стандартных ошибок, S_p, в обе стороны от выборочной доли p

Формулировка доверительного интервала для биномиального распределения (большое значение n)

В частности, мы уверены на 95%, что процент π в генеральной совокупности находится где-то между $p - tS_p$ и $p + tS_p$, где t взято из t-таблицы.

В общем случае длина доверительного интервала определяется, главным образом, размером выборки n и неопределенностью генеральной совокупности. При прочих равных условиях, если размер выборки n большой, то доверительный интервал будет меньше, поскольку формула для вычисления стандартной ошибки включает деление на $\sqrt{n}$. Уменьшение интервала свидетельствует о меньшей неопределенности из-за наличия большего количества информации. Кроме того, если для выборки неопределенность меньше, то и доверительный интервал будет меньше (это может иметь место при небольшом стандартном отклонении S или, в случае биномиального распределения, если значение доли p близко к 0 или 1).

t-таблица и t-распределение

t-таблицу, представленную здесь под заголовком *таблица 9.1.1*, часто используют для того, чтобы найти множитель при построении доверительного интервала.

[4] Отметим, что так как в данном случае n велико, значения из t-таблицы будут мало отличаться от соответствующих значений для нормального распределения. Хотя можно использовать непосредственно таблицы для нормального распределения, мы все же используем t по двум причинам: (1) поскольку вместо истинного значения стандартного отклонения σ_p используется его оценка S_p; (2) чтобы сделать подобными процедуры для выборочного среднего $\overline{X}$ и выборочной доли p, поскольку p можно считать содержательно подобным $\overline{X}$ при условии, что отдельные значения закодированы числами 0 и 1. Вычисление стандартной ошибки даст тот же результат, так как при таком кодировании $S_{\overline{X}} = S_p\sqrt{n(n-1)}$.

Таблица 9.1.1. *t*-таблица

Доверительный уровень								
Двусторонний		80%	**90%**	**95%**	98%	**99%**	99,8%	**99,9%**
Односторонний		**90%**	**95%**	97,5%	**99%**	99,5%	**99,9%**	99,95%
Уровень проверки гипотезы								
Двусторонняя		0,20	**0,10**	**0,05**	0,02	**0,01**	0,002	**0,001**
Односторонняя		**0,10**	**0,05**	0,025	**0,01**	0,005	**0,001**	0,0005
Для одной выборки: *n*	**В целом: число степеней свободы**	**Критические значения**						
2	1	3,078	6,314	12,706	31,821	63,657	318,309	636,619
3	2	1,886	2,920	4,303	6,965	9,925	22,327	31,599
4	3	1,638	2,353	3,182	4,541	5,841	10,215	12,924
5	4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
6	5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
7	6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
8	7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
9	8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
10	9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
11	10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
12	11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
13	12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
14	13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
15	14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
16	15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
17	16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
18	17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
19	18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
20	19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
21	20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
22	21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
23	22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
24	23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
25	24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
26	25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
27	26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
28	27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
29	28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
30	29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
31	30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	31	1,309	1,696	2,040	2,453	2,744	3,375	3,633
33	32	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	33	1,308	1,692	2,035	2,445	2,733	3,356	3,611
35	34	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	35	1,306	1,690	2,030	2,438	2,724	3,340	3,591
37	36	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	37	1,305	1,687	2,026	2,431	2,715	3,326	3,574
39	38	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	39	1,304	1,685	2,023	2,426	2,708	3,313	3,558
Бесконечность		1,282	1,645	1,960	2,326	2,576	3,090	3,291

Сначала обратим внимание на заголовки "Доверительный уровень" и "Двусторонний" в верхней части таблицы. Для построения обычного наиболее распространенного двустороннего 95% доверительного интервала используют соответствующую колонку таблицы. Односторонний интервал будет описан ниже в этой главе, а проверка гипотез — в следующей главе.

В статистике общее понятие **степени свободы** представляет количество независимых элементов информации, используемых для вычисления стандартной ошибки. Для одной выборки число степеней свободы равно $n-1$ (число на единицу меньше количества наблюдений), так как при вычислении стандартного отклонения из наблюдаемых значений вычитают их среднее[5]. Например, для $n=10$ наблюдений имеем 9 степеней свободы и поэтому, чтобы получить обычный двусторонний 95% доверительный интервал, используем значение 2,262 из t-таблицы. Если известно точное значение $\sigma_{\bar{x}}$, то используют значение t, равное 1,960, которое соответствует бесконечному числу степеней свободы, поскольку имеется полная информация об изменчивости. Если размер выборки n больше 40, можно в качестве допустимого приближения использовать значение t, соответствующее выборке бесконечного размера (например, 1,960 для 95% уровня доверительности). При этом ваш результат может слегка отличаться от результата компьютерных вычислений, в которых всегда используют точное значение t. Значения t из нижнего ряда таблицы (строка для выборки бесконечного размера) часто называют z-оценкой, так как они соответствуют вероятностям стандартного нормального распределения.

[5] Вычитание из наблюдаемых значений среднего приводит к потере одной степени свободы, так как сумма полученных таким образом отклонений будет равна 0 и, следовательно, только $n-1$ отклонений могут свободно изменяться, а последнее отклонение должно быть равно отрицательной сумме остальных.

Каким образом строится t-таблица? Статистики определили *t-распределение* как выборочное распределение $(\overline{X}-\mu)/S_{\overline{x}}$ при условии, что выборка взята из нормально распределенной совокупности со средним μ. (Это отношение говорит о том, на сколько стандартных ошибок, $S_{\overline{x}}$, среднее выборки $\overline{X}$ превышает среднее генеральной совокупности). Для выборки большого размера (с большим числом степеней свободы) знаменатель близок к $\sigma_{\overline{x}}$, и t-распределение становится близким к стандартному нормальному распределению. Вот почему в нижней части таблицы находятся привычные для нормального распределения значения вероятностей (такие как 1,960). Однако для выборок небольшого размера распределение отличается от нормального (рис. 9.1.4). Знаменатель $S_{\overline{x}}$ влияет на t-распределение таким образом, что кривая распределения имеет более длинный "хвост", чем при нормальном распределении. Вот почему в верхней части таблицы числа больше, чем в нижней.

Часто используемый 95% доверительный интервал

Почему чаще всего доверительные интервалы вычисляют для уровня доверительности 95%? Говорят, что именно такой выбор рассматривают традиционно как разумный. Уровень 95% представляет собой компромисс между попыткой получить по возможности более высокий уровень доверительности и желанием получить относительно небольшой интервал.

Доверительный интервал для уровня 100%, к сожалению, не очень полезен, так как он слишком велик. Представим себе следующий диалог.

Босс: "Джонс, как Вы думаете, сколько обычный потребитель будет готов платить за нашу новую марку зубной пасты?"

Джонс: "Мы считаем, что обычный потребитель будет готов платить 1,35 доллара за тюбик".

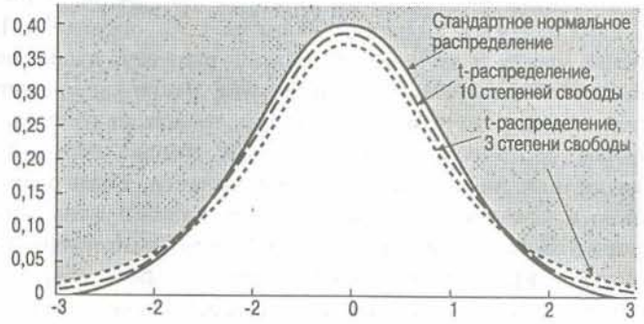


Рис. 9.1.4. t-распределение. Обратите внимание, что по мере роста размера выборки (и, следовательно, числа степеней свободы) форма кривой все больше приближается к стандартному нормальному распределению. Из-за того, что t-распределение характеризуется более длинными "хвостами", чем стандартное нормальное, приходится продвигаться дальше по кривой, чтобы захватить 95% (или исключить 5%) вероятности. Поэтому при уменьшении числа степеней свободы значения в t-таблице увеличиваются

Босс: "Насколько точна ваша оценка? На что мы можем действительно рассчитывать?"

Джонс: "Отдел анализа на 100% уверен, что типичный потребитель будет готов платить от 0 до 35 миллионов долларов за тюбик".

Подождите минутку! Это нелепо. Однако дело в том, что для полной уверенности на 100% вам необходимо рассмотреть *все*, даже самые нереальные возможности. Однако продвигаясь в сторону уменьшения от уровня 100% к доверительному уровню, который оставался бы достаточно большим, но оставлял бы некоторый простор для ошибки, можно получить реалистичный и разумный интервал. Давайте попробуем повторить этот диалог еще раз, но несколько иначе.

Босс: "Джонс, как ты думаешь, сколько обычный потребитель будет готов платить за нашу новую марку зубной пасты?"

Джонс: "Мы считаем, что обычный потребитель будет готов платить 1,35 доллара за тюбик".

Босс: "Насколько точна ваша оценка? На что мы можем действительно рассчитывать?"

Джонс: "Отдел анализа на 95% уверен, что типичный потребитель будет готов платить от 1,26 до 1,44 доллара за тюбик".

Опыт многих лет свидетельствует, что 95% доверительный уровень является удобным круглым числом, которое достаточно, но в то же время не очень близко к 100%. На практике также используют и другие доверительные уровни: 90, 99 и даже 99,9%; они будут рассмотрены после нескольких примеров.

Пример. *Контроль средней толщины бумаги*

Контрольно-измерительные приборы, установленные на оборудовании вашей бумажной фабрики, должны быть тщательно отрегулированы для того, чтобы выпускаемая бумага была стандартной толщины. Замеры толщины отобранных листов бумаги (по стандарту толщина бумаги должна равняться 0,004 дюйма) приведены в табл. 9.1.2.

Следует отметить, что средняя величина составила 0,004015 дюйма, что примерно на одну треть процента больше, чем стандартная толщина бумаги, равная 0,004 дюйма. Хотя некоторые отклонения от стандартного значения допускаются практически в любом процессе, вы должны постоянно контролировать состояние оборудования. Невозможно поверить, что среднее значение толщины, равное 0,004015 дюйма, полностью отражает результат работы оборудования. Доверительный интервал позволит вам распространить данные замеров толщины 15 отобранных листов бумаги на всю генеральную совокупность, которую можно считать либо реальной совокупностью (бумага, произведенная за текущий период), либо идеализированной (бумага, которая могла бы быть изготовлена на данном оборудовании в данных условиях). Эта идеализированная совокупность определяет текущее состояние оборудования.

При объеме выборки $n=15$ мы имеем $n-1=14$ степеней свободы и соответствующее значение из t-таблицы для двустороннего 95% доверительного интервала равно 2,145. Доверительный интервал расположен (в предположении о нормальном распределении) между точками

$$\overline{X}-tS_{\overline{x}}=0,0040146667-(2,145)(0,0000674986)=0,00387$$

и

$$\overline{X}+tS_{\overline{x}}=0,0040146667+(2,145)(0,0000674986)=0,00416.$$

Окончательное утверждение о доверительном интервале выглядит так.

"Мы на 95% уверены, что в настоящее время данное оборудование изготавливает бумагу со средней толщиной от 0,00387 до 0,00416 дюйма".

Полученный доверительный интервал показан на рис. 9.1.5. В качестве результата мы получили точное утверждение с известным уровнем доверительности об общем состоянии оборудования (или же о большом запасе бумаги, из которого была взята выборка), сделанное на основе небольшой выборки.

Таблица 9.1.2. Толщина отобранных листов бумаги (в дюймах)

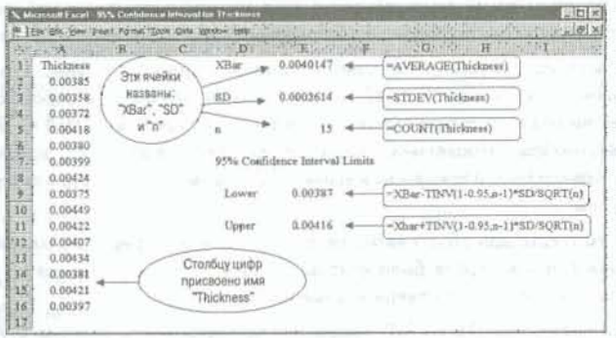
	0,00385
	0,00358
	0,00372
	0,00418
	0,00380
	0,00399
	0,00424
	0,00375
	0,00449
	0,00422
	0,00407
	0,00434
	0,00381
	0,00421
	0,00397
Среднее	0,0040146667
Стандартное отклонение	0,00026114210
Стандартная ошибка	0,0000674986
n	15



Рис. 9.1.5 Доверительный интервал для средней толщины бумаги, построенный на данных выборки объемом $n = 15$ *листов с* $\overline{X} = 0,004015$ *и* $S_{\overline{X}} = 0,0000675$ *дюйма*

Что можно предпринять, если утверждение недостаточно точно и вы хотите сузить интервал (сделать его меньше, чем от 0,00387 до 0,00416)? Для этого необходимо уменьшить стандартную ошибку $S_{\overline{X}}$[6]. Этого можно достичь двумя способами. Во-первых, уменьшить стандартную ошибку можно путем увеличения объема выборки n при прочих неизменных факторах (это легко понять, взглянув на знаменатель формулы $S_{\overline{X}} = S/\sqrt{n}$). Во-вторых, стандартная ошибка уменьшится, если даже не менять объем выборки, но уменьшить изменчивость производственного процесса (из-за числителя формулы $S_{\overline{X}} = S/\sqrt{n}$), определив важные факторы изменчивости и отрегулировав их.

Ниже приведен пример использования электронных таблиц Excel для определения доверительного интервала. Сначала дадим столбцу данных имя Thickness (Толщина), выделив его и воспользовавшись командой Insert⇨Name⇨Define (Вставка⇨Имя⇨Присвоить) из меню Excel. Затем используем функции Excel =AVERAGE(), =STDEV() и =COUNT() (=СРЗНАЧ(), =СТАНДОТКЛОН() и =СЧЁТ()) для вычисления соответственно $\overline{X}$, S и n и присвоим ячейкам соответствующие имена, чтобы к ним было легко обратиться. 95% доверительный интервал вычислим по формуле $\overline{X} \pm tS/\sqrt{n}$, используя при этом функцию Excel =TINV() (=СТЬЮДРАСПОБР()) для того, чтобы найти t-значение.[7]



Пример. *Предвыборный опрос общественного мнения (биномиальный случай)*

Президентские выборы 1992 года выглядели как гонка на трех дорожках, когда в отчете опроса *The Wall Street Journal/NBC News* сообщалось, что из 1105 опрошенных зарегистрированных избирателей 33% отдали предпочтение Пэро (независимому кандидату), 31% — Бушу (кандидату от республиканцев) и 28% — Клинтону (кандидату от демократов).[8] Такие опросы общественного мнения дают ценную информацию для людей, занимающихся организацией предвыборной кампании, а также помогают всем остальным ощущать себя в курсе текущих событий. Однако по различным причинам результаты опросов общественного мнения не являются настолько точными, как это может показаться на первый взгляд. На результаты опроса может оказать влияние то, как сформулирован вопрос (например, в каком порядке перечислены имена в списке,

[6] Можно попытаться уменьшить каким-то образом t, но это вряд ли поможет, пока исходный объем вашей выборки будет столь небольшой.

[7] Показанная в данном примере функция =TINV() получает в качестве аргумента "1 — 0,95", так как эта функция использует именно "единица минус уровень доверительности", а не сам уровень доверительности. Использовано также выражение "n–1", поскольку функция =TINV() требует в качестве аргумента количество степеней свободы.

[8] The Wall Street Journal, Western Edition, July 9, 1992, p. A1.

или то, задан ли вопрос в положительной или отрицательной форме). Предшествующие вопросы этого же интервью также могут оказывать влияние на результат, формируя некоторый сценарий и вызывая таким образом у опрашиваемого позитивное или негативное мнение. Наконец, существуют статистические ошибки выборки (возможно, как раз их легче контролировать и понять с использованием доверительных интервалов), которые указывают на то, что небольшая выборка не может служить идеальным отражением всей генеральной совокупности.

Наряду с представлением результатов опроса статья также включала следующий взгляд "за кулисы", взгляд на некоторые детали планирования и анализа результатов таких общенациональных опросов общественного мнения (текст, относящийся к доверительным интервалам, выделен курсивом).

> "В ходе опроса *The Wall Street Journal NBC News* две специализирующиеся на проведении опросов фирмы, Peter Hart и Vince Breglio, с воскресенья по среду провели по всей стране телефонные интервью с 1105 зарегистрированными избирателями.
>
> Респонденты опроса отбирались в 263 случайно отобранных географических точках континентальной части США. Каждый регион был представлен пропорционально численности его населения. Домохозяйства отбирались методом, обеспечивающим всем телефонным номерам равные возможности попасть в выборку опроса.
>
> С помощью процедуры, обеспечивающей отбор корректного числа респондентов мужского и женского пола, в каждой семье был опрошен один зарегистрированный избиратель (в возрасте 18 лет или старше). Чтобы данный опрос точно характеризовал всех зарегистрированных избирателей в общенациональном масштабе, результаты исследования были минимально взвешены по профессии респондентов.
>
> Часть вопросов была задана всем респондентам, а часть — только половине. Относительно вопросов, заданных всем, *шансы составляют 19 из 20 за то, что результаты данного опроса отличаются от результатов опроса всех имеющих телефон зарегистрированных избирателей США не более чем на три процентные единицы в любом направлении*. Предельная ошибка для вопросов, заданных только половине респондентов, составила бы 4,3 процентные единицы. Предельная ошибка для подгруппы зависела бы от размера этой группы".

В первом абзаце этого текста дан объем выборки $n = 1105$ зарегистрированных избирателей. Второй и третий абзацы поясняют, что в опросе была использована стратифицированная по регионам случайная выборка. Последний абзац описывает доверительные интервалы.

Утверждение "шансы составляют 19 из 20" указывает на вероятность, равную 0,95 или 95% (так как 19/20 = 0,95). Фраза "результатов опроса всех имеющих телефон зарегистрированных избирателей США" является четким определением *генеральной совокупности*. Слова "три процентные единицы в каждом направлении" указывают на размер доверительного интервала (tS_p), при этом речь идет о половине длины интервала, которая откладывается в обе стороны от выраженной в процентах оценки доли (p). И наконец, ссылка на предельную ошибку для подгрупп корректно указывает, что подгруппа (например, женщины или юго-западный регион) имеет меньший размер выборки, n, и поэтому здесь может иметь место большая стандартная ошибка и больший доверительный интервал, чем указанные для всей генеральной совокупности три процентные единицы.

Теперь обратите внимание на число 33%, т.е. процент опрошенных избирателей, пожелавших отдать свои голоса за независимого кандидата Пэро. Это число представляет собой точную, выраженную в процентах, долю характеристики в выборке, но в то же время это только выраженная в процентах оценка доли в генеральной совокупности. Двусторонний 95%-ный доверительный интервал находят путем прибавления и вычитания предельной ошибки к выраженной в процентах биномиальной доли. Используя значение $t = 1{,}960$ из t-таблицы, находим[9]:

$$tS_p = t\sqrt{\frac{p(1-p)}{n}} = 1{,}960\sqrt{\frac{0{,}33(1-0{,}33)}{1{,}105}} = 0{,}0277, \text{ или } 2{,}77\%.$$

[9] Здесь предельная ошибка вычисляется так, как если бы речь шла о простой случайной выборке. Для такой стратифицированной выборки, которая использована в нашем примере, необходимо произвести более сложные вычисления.

Как и было заявлено, предельная ошибка, 2,77%, действительно "не превышает три процентные единицы". 95%-ный доверительный интервал будет находиться в пределах:

от

$$p - tS_p = 0{,}33 - 0{,}0277 = 0{,}302, \text{ или } 30{,}2\%,$$

до

$$p + tS_p = 0{,}33 + 0{,}0277 = 0{,}358, \text{ или } 35{,}8\%.$$

Окончательно утверждение о доверительном интервале можно сформулировать таким образом.

> Мы на 95% уверены, что на момент проведения данного опроса от 30,2 до 35,8% имеющих телефон зарегистрированных избирателей США отдали бы свои голоса за Пэро.

Таким образом, исходя из точных данных небольшой по объему выборки мы получили точное утверждение о доверительном интервале для всех зарегистрированных избирателей США. Если быть более осторожным с выводами, то можно уточнить генеральную совокупность, включив в нее тех зарегистрированных избирателей США, до кого в период проведения опроса можно было дозвониться по телефону и кто пожелал выразить свою точку зрения. К тому же, если быть щепетильным, то следует признать, что полученные результаты свидетельствуют о том, как люди *определяют* свое отношение к Пэро, но не о том, как они будут голосовать.

Другие доверительные уровни

Хотя наиболее часто используемым доверительным уровнем является уровень 95%, иногда используют и другие доверительные уровни. Выбор уровня представляет собой поиск компромисса между размером интервала (меньший интервал является более точным, а значит, и более предпочтительным) и вероятностью того, что интервал включает искомый параметр генеральной совокупности (более высокая вероятность является более предпочтительной). В одних ситуациях необходима очень высокая точность выводов, и тогда увеличивают размеры интервала, чтобы вероятность справедливости утверждения о принадлежности параметра интервалу была выше. В других ситуациях может быть необходим более короткий интервал, и для этого можно допустить, чтобы утверждение о доверительном интервале могло быть неверным более часто. Стандартный 95% доверительный интервал является общепринятым компромиссом между этими двумя факторами, но не единственным решением данной проблемы.

При построении доверительных уровней предпочитают использовать круглые числа (избегая такие сбивающие с толку утверждения, как, например, "быть уверенным на 92,649%"). Значения t-таблицы можно использовать для построения доверительных интервалов для уровней 90, 95 и 99,9% (см. выделенные жирным шрифтом значения в верхней части таблицы), а также для других уровней, которые указаны в таблице прежде всего для построения односторонних интервалов, описанных в разделе 9.4.

Насколько меньше становится доверительный интервал при переходе к более низкому значению доверительного уровня? Для большой выборки относительные размеры доверительных интервалов показаны на рис. 9.1.6.



Рис. 9.1.6. Относительные размеры 90, 95 и 99,9% доверительных интервалов для большой выборки. Чем выше требуется уровень доверительности, тем большим должен быть размер интервала, чтобы удовлетворить заявленное требование

Пример. *Средняя продажная цена, определенная посредством скидок*

Все, наверное, слышали о торговых скидках. Они немного похожи на снижение цены при покупке товара, но для получения скидки необходимо собрать некоторые документы (например, торговый чек; этикетки от бутылок, которые нелегко отодрать, и т.п.), потратить мелкие деньги на отправку письма и, наконец, дождаться получения чека на один доллар, по которому можно получить наличные в банке.

С точки зрения производителя одно из свойств скидок заключается в том, что они дают определенную полезную информацию. Предположим, что ваша фирма имеет программу скидок на определенный вид батареек, цена которых составляет $2,39. Вам хотелось бы знать, сколько покупатели *действительно* заплатили за этот товар после снижения цен на распродажах в магазинах. Тщательно спланированное случайное исследование может дать необходимую информацию, но в данном случае такое исследование может быть не оправдано финансово. Итак, вы решили проанализировать кассовые чеки, присланные вместе с запросами на предоставление скидок.

Прежде всего, какой вид выборки мы имеем в этом случае? Это не случайная выборка в прямом смысле этого слова. Потребители, которые прислали запрос на скидки, не являются репрезентативным "срезом" всех потребителей. В частности, эти потребители могут быть лучше организованы (сохраняют все чеки для того, чтобы потом их выслать) и быть беднее (поэтому для них такая скидка — это деньги, для получения которых стоит побеспокоиться) остальных людей. Тем не менее вы решили изучить генеральную совокупность покупателей, которые высылают запросы на скидки, и рассматривать полученные вами запросы как случайную выборку из этой теоретической генеральной совокупности.

Таким образом, было получено и проанализировано $n = 15\,603$ кассовых чеков. Итоговые цифры приведены ниже.

Обобщенные данные о кассовых чеках	
n	15 603
Средняя цена продаж	$2,387
Стандартное отклонение	$0,318
Стандартная ошибка	$0,00255

Посмотрите на величину стандартной ошибки! Она так незначительна, потому что размер выборки очень большой. Ваша оценка ($2,39) очень близка к среднему генеральной совокупности.

95% доверительный интервал, рассчитанный с использованием значения $t = 1{,}960$, находится между $2,382 и $2,392. Таким образом, средняя цена в вашей идеализированной совокупности отличается от полученной оценки $2,387 не более чем на половину цента.

При такой высокой точности вы немного потеряете, сформулировав утверждение для более высокого доверительного уровня. Например, давайте используем самый высокий доверительный уровень из нашей таблицы. Чтобы достичь доверительного уровня 99,9%, подставим значение $t = 3{,}291$ вместо 1,960 и получим доверительный интервал

$$\text{от } \bar{X} - tS_{\bar{x}} = 2{,}387 - (3{,}291)(0{,}00255) = \$2{,}379$$

$$\text{до } \bar{X} + tS_{\bar{x}} = 2{,}387 + (3{,}291)(0{,}00255) = \$2{,}395\,.$$

Окончательная формулировка утверждения относительно доверительного интервала будет следующей.

> "Мы на 99,9% уверены, что средняя цена покупки батарейки потребителями, которые заинтересованы в получении скидок, находится между $2,379 и $2,395".

Сравним это с 95% доверительным интервалом. Хотя 99,9% доверительный интервал несколько больше, он все еще очень близок к оценке среднего значения цены (приблизительно на цент отличается от $2,387). Поскольку в этом случае изменчивость (измеренная стандартной ошибкой) очень низка, вы можете сделать очень точное утверждение, которое будет верным с высокой вероятностью.

Пример. *Результат производственного процесса*

Сейчас, когда введено в действие новое химико-технологическое оборудование, высшее руководство хочет получить надежную информацию о долгосрочных возможностях системы. Технологический процесс очень чувствителен, поэтому, независимо от того, насколько тщательно контролируются параметры, существует некоторое отклонение количества продукции, выпущенной в разные дни или даже в разные часы. Давайте построим доверительный интервал для объема выпускаемой продукции в долгосрочном плане (рассматривая это число как среднее генеральной совокупности), исходя из объемов продукции для выборки временных периодов.

В табл. 9.1.3 приведен столбец данных, состоящий из 12 значений объемов продукции и обычных итоговых показателей. Из таблицы видно, что данные характеризуются сильной изменчивостью.

Поскольку изменчивость данных велика, вы озабочены тем, что доверительный интервал будет больше, чем вам хотелось бы. Вы обсудили эти вопросы с другими сотрудниками и пришли к решению, что 90% доверительный интервал будет приемлемым.

Значение t для двустороннего доверительного интервала при $n - 1 = 11$ степенях свободы равно 1,796. Следовательно, доверительный интервал будет расположен между

$$\bar{X} - tS_{\bar{x}} = 60{,}3917 - (1{,}796)(5{,}4203) = 50{,}7 \text{ и}$$

$$\bar{X} + tS_{\bar{x}} = 60{,}3917 + (1{,}796)(5{,}4203) = 70{,}1\,.$$

Окончательная формулировка утверждения о доверительном интервале будет следующей:

> "Мы на 90% уверены, что рассчитанный для большого периода времени средний объем продукции для данного технологического процесса с высокой изменчивостью находится между значениями 50,7 и 70,1 тонны".

Этот 90% доверительный интервал лишь немного короче 95% доверительного интервала, который имеет границы от 48,5 до 72,3 тонны. Эффект не так уж и велик. Сделав вывод, который может быть неверным в дополнительных 5% случаев, вы получили лишь небольшое увеличение точности в сравнении с общепринятым 95% интервалом.

Таблица 9.1.3. Объем продукции химического оборудования (в тоннах)

	71,7
	46,0
	103,9
	54,4
	43,3
	68,1
	73,4
	45.1
	45,6
	44,9
	77,8
	50,5
Среднее	60,3917
Стандартное отклонение	18,7766
Стандартная ошибка	5,4203
n	12

9.2. Предположения, необходимые для корректного использования

Можно ли быть уверенным в том, что доверительные уровни дают необходимую точность? Когда заявляют о 95% уровне доверительности, можно ли быть уверенным, что с вероятностью 95% среднее генеральной совокупности действительно находится в данном интервале? Для применения статистической теории к конкретному случаю необходимо сделать некоторые технические предположения. Если эти предположения справедливы, доверительные интервалы будут определены правильно. Если предположения не справедливы для конкретной ситуации, то утверждения о доверительных интервалах могут быть неверными.

Что в действительности имеют в виду, когда говорят, что утверждение о доверительном интервале верно? Предположим, что имеется процедура для построения 95% доверительного интервала. Если эта процедура верна и ее повторяют неоднократно (т.е. строят много доверительных интервалов), то приблизительно 95% построенных с ее помощью доверительных интервалов будут содержать среднее генеральной совокупности. Таким образом, у нас нет полной гарантии, что среднее генеральной совокупности определенно находится в построенном нами интервале, хотя очень вероятно, что это среднее находится именно там.

Что имеют в виду, когда, в случае некорректности допущения, говорят, что утверждение о доверительном интервале *неверно*? Это означает, что вероятность попадания среднего генеральной совокупности в интервал *не обязательно* равна заявленному уровню доверительности 95% (или другому заявленному уровню

доверительности). Процедура построения должна давать уровень доверительности 95%, но в действительности мы можем получить намного меньший уровень доверительности, равный 50, 10% или еще меньше. Такой доверительный интервал практически бесполезен, даже если он *выглядит* как хороший. С другой стороны, если исходные предположения неверны, реальный уровень доверительности может быть даже *больше* заявленных 95%. К сожалению, в таком случае просто неизвестно, является ли действительный уровень доверительности выше или ниже заявленного уровня 95%.

Для **корректного построения доверительного интервала** необходимо выполнение двух следующих условий: (1) выборка должна быть случайной, (2) распределение должно быть нормальным. Оба этих требования должны быть удовлетворены, чтобы утверждение о доверительном интервале было корректным. Последовательно рассмотрим каждое из этих допущений.

Случайная выборка

Условие 1

Набор данных является случайной выборкой из интересующей нас генеральной совокупности.

Доверительный интервал представляет собой утверждение о среднем генеральной совокупности, основанное на выборочных данных. Естественно, должна быть сильная связь между данными и средним генеральной совокупности. Случайная выборка гарантирует, что ваши данные представляют генеральную совокупность и каждое наблюдение несет в себе новую, независимую информацию. Без случайной выборки нельзя сделать точное вероятностное утверждение о результатах. Если, например, ваша выборка состоит только из ваших друзей, то нельзя ожидать, что рассчитанный вами доверительный интервал будет отражать срез всего общества.

Можно интерпретировать это условие таким образом, что необходимо извлечь случайную выборку из тщательно определенной основы выборки, как об этом говорилось в главе 8. Конечно, в результате таких усилий требуемое условие будет выполнено. Однако это условие не является таким существенным ограничением, как это может показаться. Альтернативный способ удовлетворить необходимое условие о случайности выборки состоит в использовании идеализированной (теоретической) совокупности.

Если имеются некоторые данные и требуется построить доверительный интервал, но эти данные не представляют собой тщательно построенную случайную выборку из четко определенной генеральной совокупности, то можно попытаться построить идеализированную совокупность. Нужно задать себе вопрос о том, что представляют имеющиеся у вас данные. Если вы можете представить себе большую группу и предположить, что ваши данные могут быть рассмотрены как случайная выборка из этой большой группы, тогда обоснованным будет построение доверительного интервала, который даст вам информацию о неизвестном среднем этой идеализированной (теоретической) совокупности.[10]

[10] Однако могут быть люди, которые не согласятся с тем, как вы выделили идеализированную (теоретическую) генеральную совокупность. Поскольку это концептуальная, а не чисто статистическая проблема, здесь трудно дать рекомендации.

Например, предположим, что у вас есть данные о людях, недавно обратившихся по вопросу трудоустройства. Строго говоря, эта группа *не* является случайной выборкой из какой-либо совокупности, так как при ее отборе не использовался никакой случайный процесс. У нас недостаточно наблюдений, чтобы утверждать, что эта группа похожа на случайную выборку или что эта группа является достаточно особенной (специфичной). Таким образом, факт остается фактом — в принципе, эта группа не является случайной выборкой. Однако если вы хотите рассматривать этих людей как представителей большей совокупности лиц, ищущих работу на предприятиях, подобных вашему, то вы можете построить доверительный интервал. Этот доверительный интервал будет относиться не к тем конкретным людям, которые обратились по вопросу трудоустройства, а к похожим на них людям из идеализированной (теоретической) совокупности.

Ниже приведен пример неудачного построения доверительного интервала на основе данных, не являющихся случайной выборкой из изучаемой генеральной совокупности.

Пример. *Прогноз процентных ставок*

Экономические прогнозы позволяют уменьшить неопределенность в процессе стратегического планирования экономической деятельности. Эти прогнозы часто рассматривают как "наилучшую" доступную информацию. Может быть, это и так, хотя неизвестно, насколько в действительности надежны эти данные. Периодически *The Wall Street Journal* публикует старые прогнозы отдельных экономистов вместе с фактическими результатами, чтобы показать, насколько хорошо был сделан прогноз.

На рис. 9.2.1 показана гистограмма прогнозов ставки процента трехмесячных казначейских векселей США в конце 1998 года, которые сделали на 6 месяцев 54 экономиста.[11]

Двусторонний 95% доверительный интервал, рассчитанный на основе этих 54 прогнозов, покрывает значения ставки процента от 5,01 до 5,15% и не включает фактическое значение, равное 4,46%. Означает ли это, что нам просто не повезло (в том смысле, что в 5% случаев 95% доверительный интервал все же не включает реальное значение), или просто неразумно ожидать, что доверительный интервал может быть использован в такой ситуации? В действительности это не простая неудача; фактически в этой ситуации доверительный интервал интерпретирован неверно, поскольку этот набор данных не является выборкой из той генеральной совокупности, с которой проводится сравнение.

Рассмотрим другой набор прогнозов. На рис. 9.2.2 представлена гистограмма прогнозов ставки процента трехмесячных казначейских векселей США в конце 1984 года, которые сделали на шесть месяцев 22 экономиста.[12]

Двусторонний 95% доверительный интервал, построенный на основе 22 прогнозов, охватывает значения от 10,29 до 10,99% и также не включает фактическое значение ставки процента, равное 7,84%. Фактически прогнозы предсказания очень далеки от действительного значения.

Должен ли нас беспокоить тот факт, что доверительный интервал не включает необходимое нам значение? Необязательно, так как мы знаем, что интервалы включают необходимое нам значение только в 95% случаев. Однако вызывает удивление, что фактическое значение находится чрезвычайно далеко за пределами доверительного интервала. В нашем случае стандартная ошибка равна 0,167% и фактическое значение 7,84% находится на расстоянии $(10{,}64-7{,}84)/0{,}167=16{,}8$ стандартных ошибок от выборочного среднего (10,64%). Это расстояние, равное 16,8 стандартных ошибок, очень велико и требует объяснения.

[11] Данные взяты из C. Mitchell Ford, "U.S. Economy Is Seen Growing at Slower Pace in 1999", The Wall Street Journal, January 4, 1999, p.2.

[12] Данные взяты из T. Herman E. P. Foldessy "Interest Rates to Rise in '85 as Economy Emerges from Doldrums, According to Poll of Economists", The Wall Street Journal, 1985, January 2.

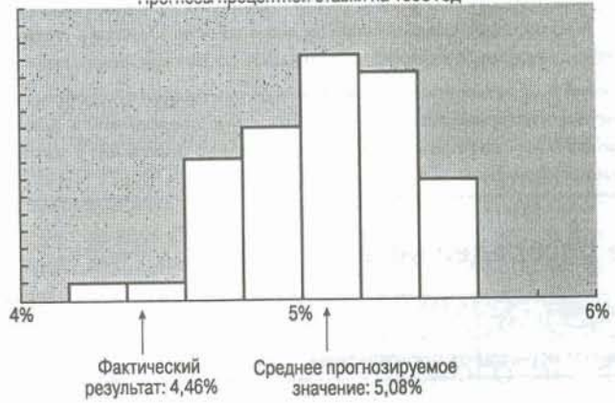


Рис. 9.2.1. Гистограмма прогнозов ставки процента трехмесячных казначейских векселей США в сравнении с фактическими ставками в конце 1998 года

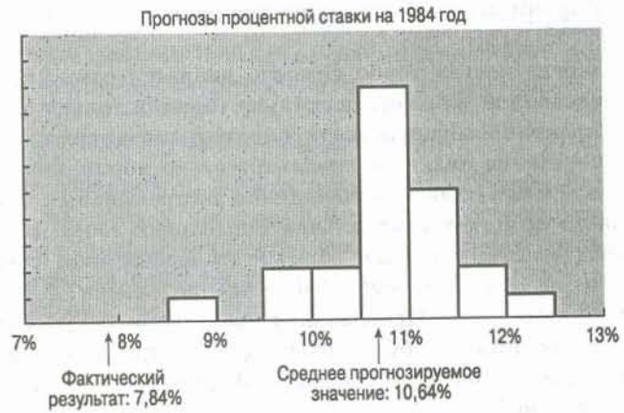


Рис. 9.2.2. Гистограмма прогнозов ставки процента трехмесячных казначейских векселей США в сравнении с фактическими ставками в конце 1984 года. В этом случае прогнозы экономистов не столь согласованы

А объяснение достаточно простое. В данном случае не выполнено условие случайности выборки и, следовательно, нет гарантии, что доверительный интервал является корректным. Таким образом, при планировании будущей деятельности следует более осторожно и критически относиться к прогнозам экономистов.

Однако что же *представляют* эти экономические прогнозы? Пожалуй, наилучшим решением будет рассматривать эти прогнозы как случайную выборку из *идеализированной генеральной совокупности* таких же прогнозов ведущих экономистов, сделанных в один и тот же период времени. Тогда такой доверительный интервал будет характеризовать согласованное среднее значение прогнозов этой группы экономистов. Однако этот доверительный интервал не будет характеризовать непосредственно ставку процента, потому что будущее значение ставки процента *не* является средним генеральной совокупности, из которой взята выборка.

Между прогнозами и фактическим значением ставки процента наблюдается большое различие. Это то, что может произойти и происходит в экономической деятельности. Однако в *статистике*, если удовлетворены необходимые исходные предположения, корректность соответствующих выводов гарантирована.

Чтобы ход размышлений был лучше понятен, следует отделить предметную область (в данном случае это экономика) от статистических принципов. Лучшее, что можно предпринять в подобной ситуации, — это сделать ограниченный точный статистический вывод и интерпретировать его в соответствии с принятыми в этой предметной области подходами.

Нормальное распределение

Условие 2

Значения данных имеют нормальное распределение.

Теория, лежащая в основе построения доверительного интервала, основана на предположении, что *в генеральной совокупности* изучаемые данные имеют нормальное распределение. Такое упрощенное условие дает возможность использовать необходимые формулы и построить t-таблицу (что уже сделано для вас). К счастью, есть две причины, по которым на практике это требование не является столь критичным.

Во-первых, никогда нельзя точно сказать, является совокупность полностью нормально распределенной или нет, поскольку имеется только случайная выборка. Поэтому на практике обращаются к гистограмме данных, чтобы удостовериться, что распределение является *приблизительно* нормальным, т.е. не слишком асимметричным и без особо отличающихся экстремальных значений.

Во-вторых, часто помогает центральная предельная теорема. Поскольку статистический вывод основан главным образом на выборочном среднем $\bar{X}$, необходимо, чтобы в первую очередь выборочное распределение $\bar{X}$ было распределено приблизительно нормально. Центральная предельная теорема утверждает, что если n достаточно велико, $\bar{X}$ будет распределено приблизительно нормально, даже если отдельные элементы в генеральной совокупности (и в выборке) не распределены нормально.

Итак, можно сформулировать следующее правило для практического применения.

Условие 2 (на практике)

Изучите гистограмму данных. Если она выглядит приблизительно нормально, то все в порядке (т.е. доверительный интервал приблизительно корректен). Если гистограмма немного асимметрична, то необходимо иметь не очень малую выборку. Если гистограмма умеренно асимметрична или имеет несколько умеренных отличающихся значений, то необходимо иметь выборку большого размера. Если гистограмма сильно асимметрична или имеет большие экстремальные значения, то это должно вызывать беспокойство.

Для биномиального распределения следствие центральной предельной теоремы заключается в том, что выраженная в процентах доля признака в выборке p при большом значении n распределена приблизительно нормально (при условии, что выраженная в процентах доля признака в генеральной совокупности не

слишком близка к 0% или 100%). Это показывает, каким образом условие нормальности распределения может быть выполнено (приблизительно) для биномиального распределения.

Как быть, если условие нормальности распределения не выполняется совсем, скажем при сильной асимметрии? Одним из способов заключается в поиске и использовании такого преобразования данных (например, логарифмического), которое привело бы к нормальному распределению; однако в таких случаях необходимо помнить, что в результате будет получен доверительный интервал для среднего *логарифмов* значений генеральной совокупности. Другой способ заключается в использовании непараметрических методов, которые будут описаны в главе 16.

9.3. Интерпретация доверительного интервала

Что вы имеете в виду, когда говорите, что исходя из значений веса в выборке из дневной продукции вы на 95% уверены, что средний вес всех изготовленных сегодня упаковок мыла лежит в пределах от 15,93 до 16,28 унций? Это похоже на вероятностное утверждение, но с ним необходимо тщательно разобраться. Средний вес всех выпущенных сегодня упаковок мыла является некоторым конкретным неизвестным числом. Это число либо принадлежит интервалу, либо не принадлежит. А раз так, то откуда появляется вероятность?

Какое событие имеет вероятность 95%?

Чтобы возникла вероятность, должен иметь место случайный эксперимент. Вероятность скорее относится к *процессу* в целом, чем к конкретному результату. Когда вы говорите, что на 95% уверены в том, что среднее значение веса в генеральной совокупности находится в пределах от 15,93 до 16,28 унций, то делаете вывод о точных числовых результатах, исходя из имеющихся данных. Однако вероятность 95% возникает из самого процесса, который рассматривает значения как *случайные*. Более тщательная формулировка вероятностного утверждения может быть такой: "Вероятность события "средний вес в генеральной совокупности находится в пределах доверительного интервала" для случайного эксперимента "случайно отобрать несколько упаковок и построить доверительный интервал" равна 95%". Каждый раз, когда собирают данные и вычисляют 95% доверительный интервал, проводят случайный эксперимент, который определяет вероятность для каждого события. Вероятность того, что неизвестное среднее генеральной совокупности попадает в вычисленный интервал, равна 0,95.

Тонкостью является вопрос о том, к какому времени относится информация. Можно обоснованно заявить, что вероятность того, что завтра индекс фондового рынка поднимется, равна 55%. Однако, когда завтра после обеда окажется, что индекс действительно пошел вверх, уже не остается ни неопределенности, ни вероятности — индекс пошел вверх. Но перед тем как это произошло, неопределенность *была*. Единственное различие между этим примером о фондовом рынке и обычным построением доверительного интервала состоит в том, что в последнем случае среднее генеральной совокупности либо находится в доверительном интервале, либо нет, однако при этом вы можете никогда не узнать, действительно ли среднее находится в этом интервале или нет!

Чтобы понять, что означает вероятность 95%, полезно представить себе многократное повторение процесса извлечения выборки и построение множества доверительных интервалов, вычисленных на данных различных случайных выборок. Из понятия относительной частоты и закона больших чисел (из главы 6) мы знаем, что приблизительно 95% этих случайных известных интервалов включают конкретное неизвестное среднее значение генеральной совокупности. Это показано на рис. 9.3.1. Обратите внимание, что каждая выборка имеет свое среднее $\bar{X}$, поэтому некоторые интервалы смещены вправо или влево относительно друг друга. Кроме того, каждый интервал имеет свою стандартную ошибку $S_{\bar{X}}$, поэтому одни интервалы больше, другие — меньше. Заметьте, что даже те доверительные интервалы, которые не содержат среднее генеральной совокупности, тем не менее, расположены к нему довольно близко.

Ваши жизненные достижения

Конечно, обычно в конкретной ситуации вы вычисляете только *один* доверительный интервал для среднего генеральной совокупности. Однако если проводится много подобных исследований, независимых друг от друга (т.е., выборки извлекаются независимо), то можно интерпретировать смысл "95% доверительности" в терминах достижений вашей жизни. Если вы за всю свою жизнь построили много 95% доверительных интервалов и если в каждом случае были удовлетворены необходимые условия, то приблизительно 95% этих доверительных интервалов действительно содержали соответствующие им средние значения генеральных совокупностей.

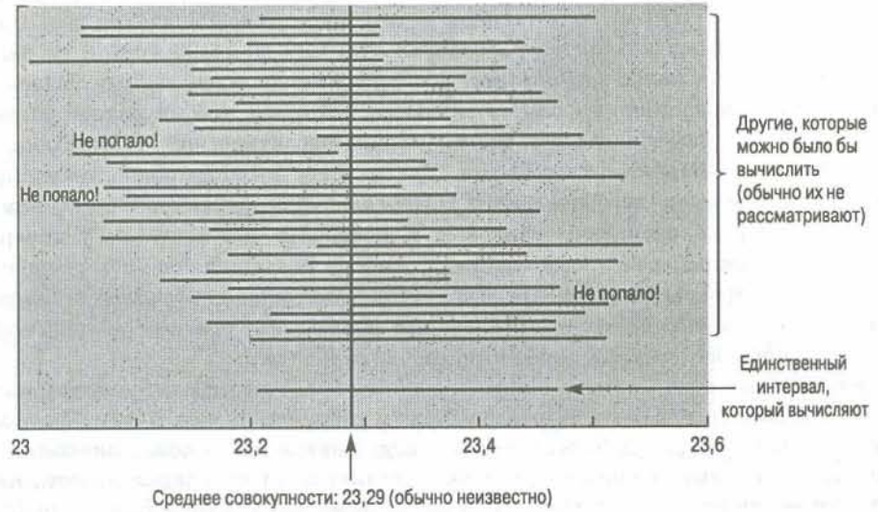


Рис. 9.3.1. Что если использовать другую случайную выборку? Этот рисунок показывает, как могут различаться доверительные интервалы, построенные на основе различных случайных выборок (построенных независимо) из одной и той же генеральной совокупности. Если процесс будет продолжаться достаточно долго, то 95% этих интервалов будет включать неизвестное среднее при условии выполнения исходных предположений

Обучаясь на курсах гольфа в доме престарелых и вспоминая свою жизнь, вы чувствуете удовлетворение от того, что в 95% случаев ваши доверительные интервалы были правильны. К этому чувству примешивается также огорчение, поскольку 5% интервалов были неверны. И вы *никогда не сможете узнать*, в каких случаях результаты вашего труда были верны, а в каких нет! Таковы методы статистического вывода.

9.4. Односторонние доверительные интервалы

В некоторых случаях нет необходимости утверждать, что среднее генеральной совокупности находится *между* двумя границами доверительного интервала. Достаточным может быть утверждение, что среднее генеральной совокупности по крайней мере *не меньше, чем* некоторое число, или (в других случаях) что среднее генеральной совокупности *не больше, чем* некоторое число. **Односторонний доверительный интервал** устанавливает с известной доверительностью, что среднее генеральной совокупности либо *не меньше*, либо *не больше* некоторого вычисленного значения. Если подходить тщательно, то построение одностороннего доверительного интервала может быть даже более эффективным, чем использование двустороннего.

Например, вас может только лишь интересовать, чтобы нечто было *достаточно большим*: "Мы на 95% уверены, что сумма продаж составит, по меньшей мере, $560 000". Или вас может только интересовать, чтобы нечто было *достаточно небольшим*: "Мы на 95% уверены, что уровень брака составляет не более, чем один дефект на 10 000 единиц производимой продукции". В ситуациях такого типа полезны односторонние доверительные интервалы.

Внимание! Не всегда можно использовать односторонний доверительный интервал

Для использования одностороннего доверительного интервала необходимо выполнение следующего важного критерия.

Критерий для использования одностороннего доверительного интервала

Чтобы использовать односторонний доверительный интервал, вы должны быть уверены, что независимо от характера данных вы будете использовать односторонний интервал с той же стороны (т.е. "не меньше чем" или "не больше чем"). Если изменение характера данных приводит к необходимости использовать односторонний интервал с другой стороны, то вы должны использовать вместо одностороннего двусторонний интервал. Если есть сомнения, используйте двусторонний интервал.

Предположим, что себестоимость единицы производимой вами продукции составляет $18, у вас есть необходимые выборочные данные и вы готовы вычислять доверительный интервал. Существует соблазн проделать это следующим образом: если оценка издержек *высокая* (больше $18), вы будете утверждать, что на 95% уверены, что издержки *по меньшей мере...*, а если оценка издержек *невысокая* (меньше $18), вы заявите, что на 95% уверены, что издержки *не более, чем... Не поддавайтесь соблазну!* Потому что в соответствии с указанным выше

критерием изменение стороны доверительного интервала с одной на другую исходя из данных выборки недопустимо. Вам следует строить двусторонний интервал (вы на 95% уверены, что издержки находятся между... и ...). Тому есть две причины. Во-первых, вас интересуют обе стороны доверительного интервала: иногда одна, иногда другая. Во-вторых, при работе с односторонним доверительным интервалом переход от одной стороны к другой может сделать недействительным ваше вероятностное утверждение и таким образом истинный уровень доверительности вашего утверждения может оказаться более низким, чем заявленный 95%.[13]

Вычисление одностороннего доверительного интервала

Чтобы вычислить односторонний интервал, сначала по табл. 9.1.1 найдите значение t, используя графу "Односторонний" для уровня доверительности в верхней части таблицы. (Используйте ту же строку таблицы, что и для двустороннего интервала, поскольку число степеней свободы не изменяется и равно $n-1$.) Например, чтобы вычислить 95% односторонний доверительный интервал для выборки размером $n=23$, следует использовать $t=1{,}717$. Для 99,9% одностороннего доверительного интервала при $n=35$ используйте $t=3{,}348$.

Далее следует выбрать *одно* из следующих утверждений для одностороннего доверительного интервала.

"Мы на 95% уверены, что среднее генеральной совокупности *не меньше чем* $\bar{X} - t_{\text{односторонний}} S_{\bar{X}}$"

или

"Мы на 95% уверены, что среднее генеральной совокупности *не больше чем* $\bar{X} + t_{\text{односторонний}} S_{\bar{X}}$."

Чтобы легче вспомнить, следует ли вычитать или прибавлять второй член, необходимо убедиться в том, что среднее значение выборки $\bar{X}$ включено в односторонний интервал. (Так должно быть, поскольку это ваша наилучшая оценка среднего генеральной совокупности.) Таким образом, если односторонний доверительный интервал направлен в сторону больших значений ("не меньше чем"), то он должен начинаться *ниже* среднего значения выборки, а если односторонний доверительный интервал направлен в сторону меньших значений ("не больше чем"), то он должен начинаться *выше* среднего значения выборки.

Рис. 9.4.1 иллюстрирует этот факт, а также сравнение одно- и двусторонних доверительных интервалов.

Начальная точка *одностороннего* 95% доверительного интервала совпадает с одной из двух граничных точек *двустороннего* 90% доверительного интервала. Дело в том, что есть две причины, по которым двусторонний интервал может оказаться неверным: среднее совокупности либо слишком велико, либо слишком

[13] В наихудшем случае при таких изменениях вы можете, заявив уровень доверительности 95%, реально получить 90% односторонний доверительный интервал. Так будет, если вы будете изменять стороны в зависимости от того, больше или меньше значение $\bar{X}$ значения μ. В таком случае у вас будет ошибка 5% в *каждом из* интервалов, и вы получите в качестве общей ошибки не 5%, на которые рассчитываете, а 10%.

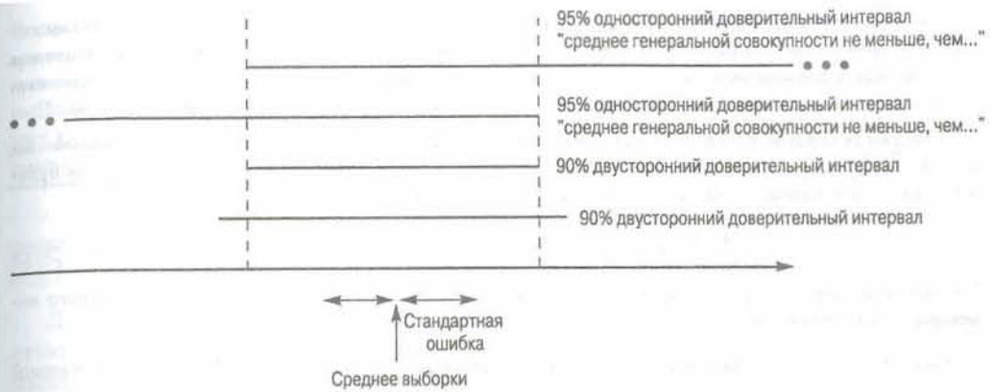


Рис. 9.4.1. В верхней части рисунка показаны оба типа односторонних доверительных интервалов. Односторонний доверительный интервал всегда включает среднее значение выборки, начинаясь с точки по одну сторону от выборочного среднего и продолжаясь до бесконечности в другую сторону. Обратите внимание, что начальная точка 95% одностороннего доверительного интервала совпадает с одной из граничных точек 90% двустороннего доверительного интервала

мало. Односторонний интервал, который имеет общую граничную точку с двусторонним интервалом, может быть неверным только в половине соответствующих случаев.

Односторонний доверительный интервал позволяет сконцентрировать внимание на наиболее интересных случаях. Если вас интересуют ошибки только по одну сторону и совсем не заботят ошибки с другой стороны, тогда односторонний интервал может иметь начальную граничную точку *ближе к среднему значению выборки* (и поэтому он будет точнее), чем двусторонний доверительный интервал. Например, для большой выборки со средним, равным 19,0, и стандартной ошибкой, равной 8,26, вместо того чтобы утверждать, что среднее генеральной совокупности лежит между 2,81 и 35,2, вы сможете сказать, что это среднее не меньше 5,41. Знание, что среднее генеральной совокупности не меньше 5,41, дает больше информации, чем знание, что это среднее не меньше 2,81. Вы можете формулировать более строгое утверждение о нижней границе благодаря тому, что вы не формулируете вообще никаких утверждений о верхней границе.

Пример. Экономия от применения новой системы

Вы оцениваете новую автоматизированную систему производства и намерены приобрести ее, если выяснится, что она обеспечивает достаточную экономию денежных средств на единицу выпускаемой продукции. Вы организовали установку системы с условием, что сможете испытать ее в течение недели. Система должна быть запрограммирована для всего ассортимента типовых изделий, и экономия будет определяться для каждой единицы выпускаемой продукции.

Что в этом случае представляет собой генеральная совокупность? Это теоретическая совокупность всех те изделий, которые могла бы выпускать система в условиях, аналогичных тем, при которых проходят испытания. Статистический вывод может помочь вам в этом случае, позволяя на основе информации о группе конкретных выпущенных изделий сделать вывод о среднем для значительно большего количества изделий, которые могли бы быть изготовлены в будущем в аналогичных условиях.

Следует ли в этом случае использовать односторонний интервал? Да, следует, поскольку независимо от поведения данных, вас интересует только одно — сможете ли вы сэкономить деньги. Интересующее вас вероятностное утверждение может выглядеть так: "Мы на 95% уверены, что среднее значение сэкономленных средств на единицу продукции при длительном периоде эксплуатации будет *не меньше* чем...".

Для выборки размером $n=18$ единиц произведенной продукции со средним значением сэкономленных средств $\bar{X}=\$39,21$ и стандартной ошибкой $S_{\bar{x}}=\$6,40$ односторонний доверительный интервал будет бесконечно простираться в направлении больших значений, начиная с

$$\bar{X}-t_{\text{односторонний}}S_{\bar{x}}=39,21-(1,740)(6,40)=28,07\ .$$

Поэтому окончательная формулировка утверждения относительно одностороннего доверительного интервала будет следующей.

> Мы на 95% уверены, что среднее значение сэкономленных средств не меньше чем $28,07 на единицу выпускаемой продукции.

Обратите внимание, что односторонний доверительный интервал включает среднее выборки $\bar{X}=\$39,21$, как и должно быть. Иными словами, выборочное среднее $\bar{X}=\$39,21$ удовлетворяет формулировке утверждения о доверительном интервале, будучи не меньше $28,07. Было бы неверно использовать другую граничную точку. Такой вид проверки правильности выбора начальной точки обеспечивает корректность построенных односторонних доверительных интервалов.

Чтобы вычислить односторонний доверительный интервал при другом уровне доверительности, просто возьмите в таблице и подставьте в формулу соответствующее t-значение. Например, при построении 99% одностороннего доверительного интервала используют значение $t=2,567$. По сравнению с 95% интервалом это утверждение относительно значения сэкономленных средств более слабое, но этому утверждению вы доверяете больше.

> Мы на 99% уверены, что среднее значение сэкономленных средств составляет не менее $22,78 на единицу выпускаемой продукции.

Пример. *Затраты на командировки*

Пытаясь спланировать реалистичный бюджет командировок, вы проанализировали стоимость типичных поездок за последнее время. Чтобы быть уверенным, что бюджет покроет требования следующего года, вам нужно получить максимальную величину средних затрат (в долларах) на одну командировку. Тогда вы сможете утверждать, что средняя стоимость командировки не будет превышать это значение. Ввиду того, что вас интересует ограничение стоимости командировки только с одной стороны, вы можете использовать односторонний доверительный интервал. Вы выбираете уровень доверительности 95%.

Анализируя список последних 83 командировок, вы определили, что средняя стоимость командировки составляла $1 286 при стандартной ошибке, равной $71,03. Односторонний доверительный интервал будет включать все значения, начиная с $0 (так как величина стоимости не может быть отрицательным числом) и до верхней границы:

$$\bar{X}+t_{\text{односторонний}}S_{\bar{x}}=1286+(1,645)(71,03)=\$1403\ .$$

Окончательная формулировка утверждения об одностороннем доверительном интервале будет выглядеть таким образом.

> Мы на 95% уверены, что среднее значение расходов на одну командировку будет не более $1403.

Чтобы удостовериться, что в использованной формуле должен стоять именно знак "плюс" (а не "минус"), посмотрим на среднее значение выборки и увидим, что это значение ($1286) действительно находится в пределах доверительного интервала ($1286 не превышает $1403).

Построенный доверительный интервал имеет ограниченное применение, поскольку использованные данные не представляют собой действительно случайную выборку из интересующей вас генеральной совокупности. Вы хотите спрогнозировать *будущие* командировочные расходы, в то время как данные вашей выборки относятся к прошлому. Доверительный интервал учитывает изменчивость командировочных затрат в прошлом, что является для вас полезной информацией. Однако он не учитывает (и не может учитывать) изменение командировочных расходов в будущем.

9.5. Интервалы предсказания

Доверительный интервал дает информацию о том, где с известной вероятностью находится *среднее генеральной совокупности*. Это очень полезно, если вы ищете обобщающую характеристику для большой генеральной совокупности. Однако если вы хотите получить информацию о наблюдаемом значении для *отдельного случая*, этот доверительный интервал вам не подходит. Вместо этого вам необходим более широкий интервал, отражающий не только оценку неопределенности $\bar{X}$, равную $S_{\bar{x}}=S/\sqrt{n}$ (которая может быть очень маленькой при большом значении n), но и оценку неопределенности S отдельного наблюдения.

Интервал предсказания позволяет использовать данные из выборки для предсказания с известной вероятностью нового наблюдения при условии, что это новое наблюдение получают тем же способом, что и прошлые данные. Таким образом, ситуация заключается в следующем. Имеется случайная выборка размером n из генеральной совокупности, и в результате проведенных для этой выборки измерений получены значения $X_1,\ldots,X_n$. Вам необходимо сделать предсказание относительно *нового* элемента, случайно выбранного из этой же генеральной совокупности.

Используемая здесь мера неопределенности представляет собой **стандартную ошибку предсказания**, т.е. величину изменчивости расстояния между средним значением выборки и новым наблюдением. Здесь объединены два вида случайности: для среднего значения выборки и для нового наблюдения. Стандартную ошибку предсказания находят умножением стандартного отклонения на корень квадратный из выражения $(1+1/n)$.

Стандартная ошибка предсказания

$$S\sqrt{1+\frac{1}{n}}$$

Значение стандартной ошибки предсказания больше оценки S изменчивости отдельных элементов в генеральной совокупности. Так и должно быть, поскольку интервал предсказания должен объединять изменчивость отдельных элементов в совокупности (которая измеряется S) и изменчивость выборочного среднего значения (которая измеряется $S_{\bar{x}}=S/\sqrt{n}$).

Вычислив оценку $(\bar{X})$ и стандартную ошибку предсказания, можно построить интервал предсказания таким же образом, как и обычный доверительный интервал. Также по таблице находят t-значение для заданного уровня доверительности прогноза и размера выборки n (конечно, не включая в выборку до-

полнительное наблюдение). Отличается только формула для вычисления стандартной ошибки; удостоверьтесь, что вы используете формулу для стандартной ошибки предсказания, а не стандартную ошибку среднего значения выборки.

Интервал предсказания для нового наблюдения

Двусторонний:

"Мы на 95% уверены, что новое наблюдение будет находиться между $\bar{X} - tS\sqrt{1+1/n}$ и $\bar{X} + tS\sqrt{1+1/n}$".

Односторонний:

"Мы на 95% уверены, что новое наблюдение не меньше, чем $\bar{X} - t_{\text{односторонний}} S\sqrt{1+1/n}$"

или

"Мы на 95% уверены, что новое наблюдение не больше, чем $\bar{X} + t_{\text{односторонний}} S\sqrt{1+1/n}$".

Какой смысл имеет здесь значение 95%? Это вероятность, соответствующая следующему случайному эксперименту: берем случайную выборку, находим интервал предсказания, берем новое случайное наблюдение и смотрим, попадает ли это новое наблюдение в построенный интервал. Обратите внимание, что вероятность 95% относится как к извлечению новой выборки, так и к извлечению нового наблюдения. Это естественно, потому что, поскольку одна выборка отличается от другой, доля новых наблюдений, попадающих в интервал предсказания, будет также разной для разных выборок. Усреднив случайность первоначальной выборки, получим вероятность, равную 95% (или другую необходимую вероятность).

В приведенной ниже таблице показано, когда необходимо использовать интервал предсказания вместо доверительного интервала.

Когда необходимо получить информацию о...	Используют ...
среднем генеральной совокупности	доверительный интервал
новом (таком же, как и остальные) наблюдении	интервал предсказания

Пример. *Сколько времени необходимо для выполнения заказа*

Когда нужно заказывать комплектующие для производства? Если вы закажете намного раньше, вам придется заплатить процент за кредит, взятый для приобретения комплектующих, и, кроме того, придется платить арендную плату за складское помещение для их хранения. Если вы закажете слишком поздно, вы рискуете остаться без необходимых комплектующих и придется на время остановить производственную линию.

При выполнении последних восьми поставок ваш поставщик говорил: "Следующий заказ будет выполнен через две недели". Вы отметили для себя, сколько рабочих дней в действительности потребовалось для выполнения последующих заказов, и у вас получились такие цифры:

$$10, 9, 7, 10, 3, 9, 12, 5$$

Среднее значение равно $\bar{X} = 8{,}125$ дня со стандартным отклонением $S = 2{,}94897$. Стандартная ошибка среднего равна $S_{\bar{X}} = 1{,}04262$, однако не она нас интересует. Стандартная ошибка прогноза вычисляется следующим образом.

Стандартная ошибка прогноза =

$$= S\sqrt{1+1/n} =$$
$$= 2{,}94897\sqrt{1+1/8} =$$
$$= 2{,}94897\sqrt{1{,}125} =$$
$$= (2{,}94897)(1{,}06066) =$$
$$= 3{,}12786$$

Для двустороннего 95% интервала предсказания t-значение из таблицы для $n = 8$ будет равно $t = 2{,}365$. Интервал предсказания находится в пределах от

$$\bar{X} - t(\text{Стандартная ошибка прогноза}) = 8{,}125 - (2{,}365)(3{,}12786) = 0{,}728$$

до

$$\bar{X} + t(\text{Стандартная ошибка прогноза}) = 8{,}125 + (2{,}365)(3{,}12786) = 15{,}52.$$

Вы предполагаете, что сроки поставки имеют приблизительно нормальное распределение. Тогда 8 наблюдений представляют собой случайную выборку из теоретической совокупности "типичных сроков поставки" и время выполнения следующего заказа на поставку является случайно выбранным из этой же генеральной совокупности. Окончательная формулировка утверждения об интервале предсказания будет следующей.

"Мы на 95% уверены, что время выполнения следующего заказа на поставку будет лежать в пределах от 0,7 до 15,5 дней".

Почему этот интервал предсказания имеет такой широкий диапазон? Диапазон отражает неопределенность ситуации. В последних 8 поставках время выполнения сильно изменялось. Естественно, это затруднило составление точного прогноза.

Если вы хотите убедиться, что следующее время поставки будет не слишком большим, вам следует построить односторонний интервал предсказания, используя $t = 1{,}895$, взятое из колонки таблицы для одностороннего 95 процентного доверительного интервала. В этом случае верхняя граница будет равна

$$\bar{X} + t_{\text{односторонний}}(\text{Стандартная ошибка прогноза}) = 8{,}125 + (1{,}895)(3{,}12786) = 14{,}05229.$$

Затем можно сформулировать следующее утверждение об одностороннем интервале предсказания.

Мы на 95% уверены, что время выполнения следующего заказа на поставку будет не больше 14 дней.

Если вас устроит 90 процентный односторонний интервал предсказания, то верхняя граница (при $t = 1{,}415$) будет следующей:

$$\bar{X} + t_{\text{односторонний}}(\text{Стандартная ошибка прогноза}) = 8{,}125 + (1{,}415)(3{,}12786) = 12{,}55092.$$

Тогда можно сформулировать следующее утверждение об одностороннем доверительном интервале.

Мы на 95% уверены, что время выполнения следующего заказа на поставку будет не больше 12,6 дней.

9.6. Дополнительный материал

Резюме

Процесс обобщения данных выборки, который приводит к вероятностным утверждениям обо всей генеральной совокупности, называют **статистическим выводом. Доверительным интервалом** называют интервал, рассчитанный из данных таким образом, что существует *известная вероятность* включения интересующего вас (неизвестного) параметра генеральной совокупности в интервал, и эта вероятность интерпретируется с точки зрения случайного эксперимента, начинающегося с извлечения случайной выборки. Вероятность того, что параметр совокупности будет принадлежать доверительному интервалу называют **уровнем доверительности**, который обычно устанавливают равным 95%, хотя часто используют и другие уровни — 90, 99, 99,9%. Чем выше уровень доверительности, тем шире (а значит, и менее полезен) доверительный интервал. Приблизительная обобщенная формулировка утверждения о доверительном интервале имеет следующий вид.

Мы уверены на 95%, что значение параметра генеральной совокупности находится между значением оценки минус две стандартные ошибки оценки и значением оценки плюс две стандартные ошибки оценки.

Это утверждение основано на том факте, что при нормальном распределении с вероятностью 0,95 следует ожидать значения на расстоянии 1,960 (приблизительно 2) стандартного отклонения от среднего.

Формулировка утверждения о двустороннем 95% доверительном интервале для среднего генеральной совокупности имеет следующий вид.

Мы уверены на 95%, что среднее генеральной совокупности μ находится между $\overline{X} - t\,S_{\overline{x}}$ и $\overline{X} + t\,S_{\overline{x}}$, где значение t берется из t-таблицы.

Для биномиального распределения (при больших n) получаем такую формулировку утверждения о доверительном интервале.

Мы уверены на 95%, что доля интересующего нас свойства в генеральной совокупности π находится между $p - tS_p$ и $p + tS_p$, где значение t берется из t-таблицы.

Чтобы получить доверительный уровень, отличный от 95%, следует просто при построении доверительного интервала использовать соответствующее значение. **t-таблицу** используют для коррекции дополнительной неопределенности, обусловленной тем, что вместо неизвестного точного значения изменчивости генеральной совокупности используют оценку (стандартную ошибку). Когда вы работаете с однократной выборкой размера n, число **степеней свободы**, равное $n-1$, представляет собой количество независимых элементов информации, использованных при вычислении стандартной ошибки (поскольку при вычислении стандартного отклонения из наблюдаемых значений вычитают среднее). Если известно точное значение стандартной ошибки, используют t-значение для бесконечного числа степеней свободы.

Для того чтобы **использование доверительного интервала было корректным**, необходимо выполнение двух следующих условий: (1) данные должны представлять собой случайную выборку из рассматриваемой генеральной совокупности,

(2) измеренные значения должны подчиняться нормальному распределению. Первое условие гарантирует, что данные правильно представляют неизвестный параметр, а второе дает основание использовать t-таблицу для вычисления вероятности. Поскольку в практической деятельности вычисление доверительного интервала основано главным образом на выборочном среднем $\overline{X}$, центральная предельная теорема позволяет смягчить условие 2, поэтому даже для умеренно асимметричного распределения при достаточно большом объеме выборки можно считать это условие выполненным.

Основанием для заявления об уверенности 95% или о доверительности 95% является то, что, как только значения для доверительного интервала вычислены, они уже не случайны; событие, имеющее вероятность 0.95, должно содержать случайность построения выборки. Интерпретация, основанная на относительной частоте, заключается в том, что если многократно повторять процесс взятия выборки и вычислять каждый раз доверительный интервал, то приблизительно 95% этих случайных известных интервалов будут включать конкретное неизвестное среднее значение генеральной совокупности. Подобным же образом из вычисленных вами в течение всей жизни при выполнении необходимых условий доверительных интервалов приблизительно 95% будут правильными (т.е. будут содержать неизвестный параметр) и приблизительно 5% будут ошибочными. Однако, вообще говоря, вы не будете знать, какие из них верны, а какие нет.

Односторонний доверительный интервал с известной доверительностью указывает, что среднее генеральной совокупности либо *не меньше*, либо *не больше* некоторого вычисленного значения. Вы вычисляете граничное значение для одностороннего доверительного интервала таким же образом, как и для двустороннего интервала, только заменяете t-значение для двустороннего интервала на t-значение для одностороннего интервала и выбираете граничную точку интервала так, чтобы построенный односторонний интервал включал выборочное среднее $\overline{X}$. При использовании одностороннего интервала вы должны быть уверены, что независимо от *поведения данных* вы будете использовать односторонний интервал с той же стороны (т.е. открытый в сторону больших значений или открытый в сторону меньших значений). В противном случае использование одностороннего доверительного интервала некорректно. При наличии сомнений лучше использовать двусторонний интервал. Утверждение об одностороннем доверительном интервале формулируется следующим образом.

Мы уверены на 95%, что среднее генеральной совокупности *не меньше*, чем $\overline{X} - t_{\text{односторонний}} S_{\overline{X}}$

или

Мы уверены на 95%, что среднее генеральной совокупности *не больше*, чем $\overline{X} + t_{\text{односторонний}} S_{\overline{X}}$.

Интервал предсказания позволяет использовать данные выборки для предсказания с известной вероятностью значения нового наблюдения при условии, что это новое наблюдение получено тем же способом, что и предшествующие. В качестве меры неопределенности здесь используется **стандартная ошибка предсказания** $S\sqrt{1+1/n}$, мера изменчивости расстояния между средним значением выборки и новым наблюдением. Интервал предсказания строят тем же способом, что и доверительный интервал; просто заменяют стандартную ошибку среднего

на стандартную ошибку предсказания. Формулировка утверждения об интервале предсказания (двустороннем) для значения нового наблюдения будет следующей.

Мы уверены на 95%, что новое наблюдение будет находиться между $\bar{X} - tS\sqrt{1+1/n}$ и $\bar{X} + tS\sqrt{1+1/n}$.

Формулировка утверждения об интервале предсказания (одностороннем) для значения нового наблюдения будет такой.

Мы уверены на 95%, что новое наблюдение будет не меньше, чем $\bar{X} - t_{\text{односторонний}} S\sqrt{1+1/n}$

или

Мы уверены на 95%, что новое наблюдение будет не больше, чем $\bar{X} + t_{\text{односторонний}} S\sqrt{1+1/n}$.

Выбирая соответствующие t-значения из таблицы, можно строить интервалы предсказания для уровней доверительности, отличных от 95%. Необходимо помнить, что доверительный интервал дает информацию о среднем генеральной совокупности, в то время как интервал предсказания дает информацию о единственном новом наблюдении, случайно выбранном из той же генеральной совокупности.

Основные термины

- Статистический вывод (statistical inference), 397
- Доверительный интервал (confidence interval), 397
- Уровень доверительности (confidence level), 397
- t-таблица (t-table), 401
- Степени свободы (degrees of freedom), 403
- Необходимые для построения доверительного интервала предварительные условия (assumptions required for the confident interval), 413
- Односторонний доверительный интервал (one-sided confidence interval), 419
- Интервал предсказания (prediction interval), 423
- Стандартная ошибка предсказания (standard error for prediction), 423

Контрольные вопросы

1. В чем статистический вывод выходит за пределы простого обобщения данных?
2. Какую дополнительную информацию о генеральной совокупности дает доверительный интервал по сравнению со значением оценки параметра?
3. Какое свойство нормального распределения приводит к тому, что в обобщенной формулировке утверждения о доверительном интервале фигурирует число 2 (или 1,96)?
4. Почему правильно говорить: "Мы на 95% уверены, что среднее совокупности находится между $15,85 и $19,36", а не "Вероятность того, что среднее совокупности находится между $15,85 и $19,36, составляет 0,95"?
5. Почему для двустороннего 95%-ного доверительного интервала t - таблица содержит значения, большие чем 1,960?

6. а) Чему равно число степеней свободы для однократной выборки размером n?

   б) Почему при вычислении числа степеней свободы вычитают единицу?

   в) Как определяется число степеней свободы, если известно точное значение стандартной ошибки?
7. а) Какие еще уровни доверительности, кроме 95%, используют достаточно часто?

   б) Чем отличается вычисление 99% доверительного интервала от 95% интервала?

   в) Какой из двух доверительных интервалов больше: двусторонний 99% или двусторонний 95%?
8. а) Укажите два условия, которые должны выполняться для корректного построения доверительного интервала.

   б) Для каждого из условий приведите пример неверного результата в случае нарушения этого условия.

   в) Каким образом центральная предельная теорема помогает выполнению одного из этих условий?

   г) При каких обстоятельствах центральная предельная теорема не гарантирует выполнения второго условия?
9. а) Дайте основанную на относительной частоте интерпретацию корректности доверительного интервала.

   б) Дайте интерпретацию корректности большого количества доверительных интервалов с точки зрения "жизненных достижений".
10. а) Почему односторонний доверительный интервал всегда должен включать среднее выборки?

    б) Должен ли односторонний доверительный интервал всегда включать в себя среднее генеральной совокупности?
11. а) Какой дополнительный критерий должен удовлетворяться для корректности одностороннего доверительного интервала (в дополнение к двум условиям для двустороннего доверительного интервала)?

    б) При наличии сомнений, какой доверительный интервал вы будете использовать: односторонний или двусторонний?
12. а) Какова разница между интервалом предсказания и доверительным интервалом?

    б) Какой тип интервала вы использовали бы, чтобы узнать о среднем значении привычных расходов вашего типичного покупателя?

    в) Какой тип интервала вы использовали бы, чтобы узнать о среднем значении привычных расходов отдельного покупателя?
13. а) Как вычисляют стандартную ошибку предсказания?

    б) Почему стандартная ошибка прогноза больше стандартного отклонения S?

14. а) Что нужно изменить в вычислениях двустороннего 95% интервала предсказания, чтобы получить вместо него двусторонний 99% интервал предсказания?

б) Что нужно изменить в вычислениях двустороннего 95% интервала предсказания, чтобы получить вместо него односторонний 95% интервал предсказания?

в) Что нужно изменить в вычислениях двустороннего 95% интервала предсказания, чтобы получить вместо него односторонний 90% интервал предсказания?

Задачи

1. Ваша сельскохозяйственная фирма собирается приобрести некоторый большой участок пригодной для обработки земли. Для принятия решения необходимо изучить плодородие земли на этом участке. Случайная выборка из 62 небольших участков демонстрирует среднюю урожайность 103,6 бушелей кукурузы с акра со стандартным отклонением 9,4 бушеля с акра. Постройте двусторонний 95% доверительный интервал для среднего урожая, собранного со всего большого участка земли, возможность покупки которого изучается.
2. Ваша компания производит и распространяет замороженные пищевые продукты. Одна упаковка данного продукта должна иметь вес — 24,5 унции. Была взвешена случайная выборка из дневной продукции, и результаты оказались следующими: средний вес 24,41 унции, стандартное отклонение — 0,11 унции, размер выборки — 5 упаковок. Постройте двусторонний 95% доверительный интервал для среднего веса всех упаковок, выпущенных за этот день.
3. Ваша больница ведет переговоры с фирмой, занимающейся медицинским страхованием. Эта фирма намерена уменьшить суммы оплаты пребывания в больнице. Для конкретного вида лечения страховая фирма хотела бы уменьшить сумму оплаты пребывания в больнице на $300, а также снизить время пребывания пациентов в больнице на один день. Чтобы определить влияние этих действий на больничные расходы, проанализировали случайную выборку из 50 пациентов, недавно прошедших такой курс лечения. В случае их выписки из больницы на день раньше средняя экономия составила бы $322,44 со стандартным отклонением $21,71. Постройте двусторонний 95% доверительный интервал для среднего размера экономии на одного пациента для большей совокупности недавно прошедших лечение пациентов.
4. Ваш отдел технического контроля качества проанализировал содержание 20 случайно отобранных бочек с материалами, которые используют при изготовлении пластикового садового инструмента. Получены следующие результаты: среднее — 41,93 галлона пригодного к употреблению материала в каждой бочке со стандартной ошибкой 0,040 галлона на бочку. Определите двусторонний 95% доверительный интервал для среднего генеральной совокупности.

5. Замерена интенсивность светового потока восьми карманных фонариков. Определите соответствующие t-значения из t-таблицы для вычисления следующих доверительных интервалов.

а) Двусторонний, доверительность 95%.

б) Двусторонний, доверительность 99%.

в) Двусторонний, доверительность 99,9%.

г) Двусторонний, доверительность 90%.

6. Проведены исследования издержек для 21 производственной ситуации. Определите соответствующие t-значения из t-таблицы для вычисления следующих доверительных интервалов.

а) Двусторонний, доверительность 95%.

б) Двусторонний, доверительность 99%.

в) Двусторонний, доверительность 99,9%.

г) Двусторонний, доверительность 90%.

7. Получены данные о реакции на вакцинацию у 1859 человек. Определите соответствующие t-значения из t-таблицы для вычисления следующих доверительных интервалов.

а) Двусторонний, доверительность 95%.

б) Двусторонний, доверительность 99%.

в) Двусторонний, доверительность 99,9%.

г) Двусторонний, доверительность 90%.

8. Сорок восемь посетителей ресторана поставили оценки первому блюду. Определите соответствующие t-значения из t-таблицы для вычисления следующих доверительных интервалов.

а) Двусторонний, доверительность 95%.

б) Двусторонний, доверительность 99%.

в) Двусторонний, доверительность 99,9%.

г) Двусторонний, доверительность 90%.

9. Собранные данные об урожайности имеют 17 степеней свободы. Определите соответствующие t-значения из t-таблицы для вычисления следующих доверительных интервалов.

а) Двусторонний, доверительность 95%.

б) Двусторонний, доверительность 99%.

в) Двусторонний, доверительность 99,9%.

г) Двусторонний, доверительность 90%.

10. Исследованы предпочтения потребителей в ситуации, для которой известна стандартная ошибка. Определите соответствующие t-значения из t-таблицы для вычисления следующих доверительных интервалов.

а) Двусторонний, доверительность 95%.

б) Двусторонний, доверительность 99%.

в) Двусторонний, доверительность 99,9%.

г) Двусторонний, доверительность 90%.

д) Односторонний, доверительность 95%.

е) Односторонний, доверительность 99%.

11. Опрошена случайная выборка из восьми покупателей с целью определения количества персональных компьютеров, которые они планируют заказать в следующем году. Получены следующие результаты: 22, 18, 24, 47, 64, 32, 45 и 35. Вас интересует информация о большей генеральной совокупности, которую представляют опрошенные покупатели.

    а) Найдите обычную обобщенную меру изменчивости для отдельных покупателей.

    б) Насколько приблизительно отличается значение среднего выборки от среднего генеральной совокупности?

    в) Определите 95% доверительный интервал для среднего генеральной совокупности.

    г) Определите 99% доверительный интервал для среднего генеральной совокупности.

12. Ваша пекарня выпекает батоны хлеба, вес которых, как указано на этикетке, составляет 1 фунт. Ниже приведены веса батонов, случайно отобранных из сегодняшней выпечки.

    1,02; 0,97; 0,98; 1,10; 1,00; 1,02; 0,98; 1,03; 1,05; 1,02; 1,06.

    Определите 95% доверительный интервал для среднего веса всех батонов из сегодняшней выпечки.

13. Маркетинговое исследование на основе выборочного опроса 483 человек показало, что в следующем году один человек будет тратить на ваше изделие в среднем $15,48. Стандартное отклонение выборки составило $2,52. Определите двусторонний 95% доверительный интервал для средних трат одного человека в будущем году для большей генеральной совокупности.

14. Последний опрос 252 потребителей, случайно отобранных из базы данных, содержащей 12 861 потребителей, показал, что 208 из них удовлетворены уровнем сервиса. Определите 99% доверительный интервал для процента удовлетворенных сервисом людей во всей базе данных.

15. Анализ выборки размером 258 человек, случайно отобранных из населения города численностью 750 339 человек, показал, что 165 из них поддерживают новый проект общественных работ. Определите 99,9% доверительный интервал для процента поддерживающих проект людей во всем городе.

16. Из 763 случайно отобранных человек 152 не знакомы с вашей продукцией.

    а) Оцените процент людей в генеральной совокупности (из которой была извлечена эта выборка), не знакомых с вашей продукцией.

    б) Определите стандартную ошибку оценки, вычисленной в п. "а".

в) Определите двусторонний 95% доверительный интервал для процента таких людей в генеральной совокупности.

г) Определите односторонний 99% доверительный интервал, согласно которому процент таких людей в совокупности не превышает определенное значение.

д) Почему этот статистический вывод является приблизительно корректным, даже если распределение в генеральной совокупности не является нормальным?

17. При проведении общенационального опроса требуется, чтобы границы ошибки не превышали 3 процентные единицы в каждом направлении (т.е. плюс или минус) при уровне доверительности 95%.

    а) Проверьте выполнение этого требования для конкретного случая, вычислив произведение t-значения из таблицы на стандартную ошибку биномиально распределенной доли ρ для ситуации, когда 309 из 1105 зарегистрированных избирателей утверждают, что они поддержат на выборах определенного кандидата.

    б) Постройте 95% доверительный интервал для выраженной в процентах доли зарегистрированных избирателей, которые поддержат определенного кандидата, как это указано в п. "а".

18. При проведении общенационального опроса граница ошибки не должна превышать 4,3 процентные единицы для вопросов, заданных половине выборки.

    а) Проверьте выполнение этого требования для конкретного случая, вычислив произведение t-значения из таблицы на стандартную ошибку биномиально распределенной доли ρ для случая, когда 46% из 553 зарегистрированных избирателей-женщин утверждают, что они поддержат на выборах определенного кандидата.

    б) Постройте 95% доверительный интервал для выраженной в процентах доли зарегистрированных избирателей-женщин, которые поддержат определенного кандидата, как это указано в п. "а".

19. В результате опроса 15 инженеров, работающих в отраслях металлообрабатывающей промышленности, которые пребывают в своей должности от одного до трех лет и являются членами Американского общества контроля качества продукции, было установлено, что их средняя зарплата составляет $37 496 в год.[14] Будем считать, что это случайная выборка со стандартным отклонением $9000.

    а) Определите 95% доверительный интервал для средней зарплаты в генеральной совокупности.

    б) Определите 99% доверительный интервал для средней зарплаты в генеральной совокупности.

    в) Закончите следующее предложение: "Мы на 95% уверены, что средняя зарплата в генеральной совокупности по крайней мере составляет _____".

[14] Quality Progress, November 1995, p. 58.

20. В результате опроса девяти менеджеров транспортной и аэрокосмической отрасли, которые работают в своей должности более 20 лет и являются членами Американского общества контроля качества продукции, было установлено, что их средняя зарплата составляет $67 333 в год.[15] Будем считать, что это случайная выборка со стандартным отклонением $19 000.

а) Определите 95% доверительный интервал для средней зарплаты в генеральной совокупности.

б) Определите 99% доверительный интервал для средней зарплаты в генеральной совокупности.

в) Закончите следующее предложение: "Мы на 95% уверены, что средняя зарплата в генеральной совокупности по крайней мере составляет _____".

21. Ваша фирма хочет нанять опытного квалифицированного менеджера по контролю качества продукции. Считая, что такой менеджер отобран случайно из генеральной совокупности, представленной в предыдущей задаче, закончите следующее предложение: "Мы на 99% уверены, что зарплата выбранного менеджера должна составлять по крайней мере _____".

22. Многие потребители считают, что они смогут экономить деньги, делая покупки в новом магазине под названием SuperMall. В табл. 9.6.1 показаны результаты исследования случайно отобранных товаров. Определите 95% доверительный интервал для среднего значения (в долларах) сэкономленных средств в генеральной совокупности, которую представляют эти товары.

23. Исходя из приведенных ниже дневных процентных колебаний индекса S&P 500 фондового рынка в июле 1995 года определите доверительный интервал для среднего дневного изменения в генеральной совокупности. (Это, строго говоря, не является случайной выборкой из генеральной совокупности. Однако теория случайных изменений фондового рынка предполагает, что колебания рынка должны вести себя как случайная выборка. Генеральная совокупность будет представлять собой все дневные изменения на рынке, которые могли бы произойти в условиях, преобладающих в данное время.)

0,43%; 0,03%; 1,23%; 0,43%; 0,15%; −0,43%; 1,10%; 0,02%;
−0,20%; 0,51%; −0,76%; −1,34%; 0,46%; 0,01%; 0,54%; 0,80%;
0,09%; 0,64%; −0,41%; −0,15%

24. Во время однонедельного эксперимента по всей стране в магазинах вашей фирмы был изменен способ демонстрации рекламы товара внутри магазина. В результате объем продаж для этих товаров увеличился (по сравнению с предыдущей неделей) в среднем на $441,84 со стандартным отклонением $68,91. В эксперименте участвовало 18 магазинов. (Аналогичные ситуации были исследованы компанией Bennett-Chaikin, Inc., о чем сообщалось в рекламе корпорации Menasha Corporation в *Advertising Age*, August 21, 1995, p. 17.)

[15] Quality Progress, November 1995, p. 61.

Таблица 9.6.1. Цены в магазине SuperMall и в других местах на различные товары

Товар	Цена в SuperMall, дол.	Цена в других местах, дол.	Экономия при покупке в SuperMall, дол.
Levi's	27,99	37,99	10,00
Reebok Walk Leader	64,99	69,99	5,00
Farberware Omelette Pan	89,99	89,99	0,00
Chocolate Meal Replacement Shake	12,99	12,99	0,00
Kolcraft Baby Stroller	109,75	119,99	10,24
Wilson's Lambskin Jacket	199,00	199,00	0,00
Joop!	47,95	55,00	7,05
Eureka Tent	239,99	239,99	0,00
Geoffey Beene Shirt	24,99	39,50	14,51
JVC VCR	398,87	386,97	−11,90
Nike Basketball Shoes	79,99	79,99	0,00
Spalding NBA Ball	69,99	69,99	0,00
Smoothe as Silk	12,99	12,99	0,00
SBA Coffee	9,50	8,95	−0,55
Barbie House	64,99	64,99	0,00
"Shameless" CD	15,99	15,99	0,00

Данные взяты из *The Seattle Times*, September 22, 1995, p. A16.

а) Определите 95% доверительный интервал для среднего значения увеличения продаж в генеральной совокупности.

б) Закончите следующее предложение: "Мы на 95% уверены, что среднее значение объема продаж в генеральной совокупности увеличится не меньше чем на _____".

в) Менеджер одного из ваших магазинов хотел бы оценить возможное увеличение продаж. Этот магазин не вошел в исследование. Предполагая, что условия в этом магазине аналогичны условиям в тех магазинах, которые принимали участие в эксперименте, закончите следующее предложение: "Мы на 95% уверены, что при изменении способа демонстрации рекламы увеличение недельного объема продаж в данном магазине составит от ___ до ___".

25. В табл. 9.6.2 представлены значения доходности акций, которые предлагают брокерские фирмы.

а) Вычислите среднее и дайте краткую интерпретацию найденному значению.

б) Вычислите стандартное отклонение и дайте краткую интерпретацию найденному значению.

в) Вычислите стандартную ошибку и дайте краткую интерпретацию найденному значению.

Таблица 9.6.2. Доходности акций, предложенных брокерскими фирмами

Фирма	Доходность, %	Фирма	Доходность, %
Thomson McKinnon	15,8	A. G. Edwards	1,2
Shearson Lehman	5,5	PaineWebber	0,4
Prudential-Bache	4,7	E. F. Hutton	0,3
Merrill Lynch	2,9	Dean Witter	–2,8
Drexel Burnham	2,5	Kidder Peabody	–7,1

Взято из J. R. Dorfman "Do Brokers' Stock Picks Perform Well?", *The Wall Street Journal*, February 5, 1988, p. 17.

г) Определите двусторонний 95% доверительный интервал среднего значения доходности акций, предложенных аналогичными брокерскими фирмами в течение этого же периода времени, рассматривая имеющийся набор данных как случайную выборку из этой идеализированной совокупности.

д) Определите двусторонний 90% доверительный интервал и сравните его с 95% доверительным интервалом.

е) Найдите односторонний 99% доверительный интервал для случая, когда среднее значение доходности *не меньше* чем некоторое значение.

ж) Предположим, анализируя данные, вы видите, что идет уменьшение среднего значения доходности, и хотите сформулировать утверждение об одностороннем доверительном интервале о том, что среднее значение доходности *не больше* чем некоторая величина (вместо вашего ответа в п. "е"). В этом случае и при использовании одной и той же таблицы данных будет ли сформулированное вами в п. е утверждение о доверительном интервале корректным? Почему да или почему нет? Если нет, то что нужно делать вместо этого?

26. Предвыборный опрос 921 случайно отобранного избирателя показал, что ваш кандидат впереди с 52,4% голосов.

а) Определите двусторонний 95% доверительный интервал для процента голосов за вашего кандидата в генеральной совокупности.

б) Поскольку вы давно симпатизируете этому кандидату и желаете ей победы на выборах, вас интересует только информация о том, будет ли за нее отдано не менее некоторого процента голосов избирателей. Будет ли правильным в данном случае строить односторонний доверительный интервал?

в) С учетом информации из п. "б" постройте подходящий односторонний 95% доверительный интервал.

г) Постройте подходящий односторонний 90% доверительный интервал.

д) Опишите в одном абзаце текста, как эти доверительные интервалы проливают свет на шансы вашего кандидата. В частности, насколько больше вы знаете теперь в сравнении с одним числом 52,4%?

27. Исследование рынка, основанное на выборочном опросе 400 человек, показало, что люди готовы тратить в следующем году на покупку вашей продукции в среднем $2,34 (на одного человека). Стандартное отклонение выборки составило $0,72. Определите двусторонний 95% доверительный интервал для средних трат одного человека в будущем году для большей генеральной совокупности.

28. К вашему удивлению, опрос ваших клиентов показал, что из случайно отобранных 200 клиентов 42 недовольны послепродажной поддержкой и обслуживанием.

а) Вычислите основные статистики: размер выборки, n; выборочный процент, p; стандартную ошибку, S_p.

б) Постройте двусторонний 95% доверительный интервал для процента недовольных среди *всех* ваших клиентов (т.е. не только тех, кто был опрошен).

в) Ваша генеральная совокупность состоит из 28 209 клиентов. Преобразуйте проценты, представляющие граничные точки доверительного интервала в п. "б", в количество людей в генеральной совокупности. Сформулируйте и объясните ваш результат как доверительный интервал для *количества* недовольных потребителей в генеральной совокупности.

29. Выборка 93 рулонов листовой стали показала, что средняя длина рулона составляет 101,37 метра со стандартным отклонением 2,67 метра.

а) Дайте словесную интерпретацию стандартного отклонения; в частности, каким образом стандартное отклонение измеряет изменчивость?

б) Найдите стандартную ошибку. Дайте словесную интерпретацию найденного значения и сравните ее с представленной в п. "а" интерпретацией стандартного отклонения.

в) Определите двусторонний 95% доверительный интервал для средней длины рулона стали в большей генеральной совокупности. Кратко опишите полученное значение.

г) Определите двусторонний 95% интервал предсказания для длины рулона, который будет произведен следующим. Кратко опишите полученный результат и сравните этот интервал предсказания с доверительным интервалом из п. "в". В частности, почему интервал предсказания настолько шире доверительного интервала?

д) Ваша совесть требует, чтобы рулоны стали были гарантированно *не меньше* определенной длины. Заставит ли эта информация вас вычислять односторонний интервал? Почему да или почему нет?

е) Определите соответствующий односторонний 99% доверительный интервал и объясните полученный результат.

ж) Определите соответствующий односторонний 99% интервал предсказания и объясните полученный результат.

30. В процессе написания брошюры, описывающей быстродействие новой компьютерной системы, вы измеряете, сколько времени необходимо специальной тестовой программе для обработки некоторой базы данных. Ввиду того что по мере добавления, удаления или изменения записей физическое раз-

мещение базы данных на диске постоянно меняется, в результатах выполнения измерительного теста наблюдается некоторая вариация. Ниже приведены значения времени (в минутах) для 14 независимых повторений теста.

5, 6, 8, 11, 5, 8, 11, 10, 6, 10, 5, 9, 5, 5

а) Определите двусторонний 95% доверительный интервал и опишите полученное значение с точки зрения среднего значения производительности системы при ее массовом выпуске.

б) Определите подходящий односторонний 95% доверительный интервал исходя из того, что вы хотите показать, насколько быстро работает система (лучшими являются более низкие значения).

в) Определите соответствующий односторонний 90% доверительный интервал.

г) Определите соответствующий односторонний 99% доверительный интервал.

д) Напишите краткий отрывок для рекламной брошюры, описывающий один (или несколько) из полученных результатов. Будьте честны, но "покажите товар лицом" и пишите обычным доступным языком. Включите в качестве сносок дополнительную техническую информацию, чтобы более подготовленные читатели могли понять подробности проделанной вами работы.

31. Проведен анализ выборки горной породы, взятой из различных мест в некоторой шахте. Для каждой выборки вычислено значение "нормы прибыли", представляющей собой прибыль (полученную после продажи выплавленного металла по текущей рыночной цене) как процент от стоимости добычи. Эта величина отражает затраты на добычу руды, ее переработку и доходность конечного продукта, все в конкретных экономических условиях. Можно рассматривать эти данные как случайную выборку, которая соответствует реальным условиям функционирования экономически жизнеспособного производства. Для экономической устойчивости необходимо, чтобы норма прибыли была достаточно высокой и оправдывала затраты на производство. Ниже приведены значения нормы прибыли, полученные для 13 выборок.

8,1%; 6,2%; 19,8%; –4,3%; 5,1%; 0,2%; –10,4%; 11,8%; 2,0%; 4,7%; –3,2%; 8,9%; –6,2%

а) Вычислите основные статистики: n, среднее, стандартное отклонение и стандартную ошибку. Кратко опишите ситуацию, представив, что вы должны дать объяснение совету директоров.

б) Определите генеральную совокупность и среднее генеральной совокупности. Почему значение среднего генеральной совокупности является важным как для управления, так и для владельцев данной шахты?

в) Определите соответствующий односторонний 99% доверительный интервал. Письменно поясните полученное значение.

г) Подготовьте краткую докладную записку для высшего руководства, в которой опишите в общих чертах ситуацию и дайте рекомендации возможных действий.

32. Вас беспокоит брак при печати газеты. Никакие количественные измерения ранее не проводились, несмотря на то, что частые ошибки вынуждают выбрасывать много фунтов только что напечатанной продукции. Чтобы оценить серьезность проблемы и помочь в ее решении, если это будет оправдано, вы начали собирать данные. Вы будете предпринимать какие-то меры только в том случае, если количество брака достаточно велико. Взвешены бракованные экземпляры газеты для 27 отобранных утренних выпусков. Средний вес брака составил 273,1 фунта в день со стандартным отклонением 64,2 фунта.

а) Подходит ли для данной ситуации построение одностороннего доверительного интервала? Почему да или почему нет?

б) Постройте наиболее полезный в данной ситуации односторонний 99% доверительный интервал. Почему вы выбрали именно такой односторонний интервал?

в) Выразите построенный доверительный интервал в фунтах за год, приняв, что в году 365 рабочих дней.

г) Определите односторонний 99% интервал предсказания для завтрашнего количества брака. Сравните полученный результат с результатом из предыдущего пункта.

33. На новой работе вы заключили девять контрактов на продажу продукции со средней ценой $3782 и стандартным отклонением $1290.

а) Определите подходящую идеализированную совокупность, которую представляет данная выборка.

б) Если распределение цен продаж достаточно сильно асимметрично, можно ли строить обычный двусторонний 95% доверительный интервал? Почему да или почему нет?

в) Теперь предположим, что распределение цен только слегка асимметрично и не сильно отличается от нормального. Вычислите обычный двусторонний 95% доверительный интервал и поясните его с точки зрения долгосрочной перспективы вашей работы на этом месте. Отразите как полезную информацию, представленную доверительным интервалом, так и ограничения доверительного интервала.

г) Найдите двусторонний 90% интервал предсказания для цены продаж следующего заключенного вами контракта, приняв, что условия останутся в основном теми же.

34. Анализ случайной выборки записей о 50 пациентах, недавно посетивших клинику, демонстрирует, что средняя стоимость одного визита к врачу составляет $53,01 со стандартным отклонением $16,48.

а) Найдите 95% доверительный интервал для среднего и дайте его интерпретацию.

б) Найдите 99% доверительный интервал.

в) Найдите односторонний 95% доверительный интервал, определяющий по крайней мере некоторый уровень стоимости визита к врачу.

35. Найдите 95% доверительный интервал для суммы денег, которую ваши постоянные покупатели потратили на ваши товары в прошлом месяце, используя данные из задачи 2 главы 4 в качестве случайной выборки заказов покупателей.

36. Найдите 99% доверительный интервал для прочности хлопковой нити, используемой на ткацкой фабрике, используя данные задачи 15 главы 4.

37. Определите 99,9% доверительный интервал для веса шоколадного батончика до вмешательства в процесс производства, используя данные задачи 11 главы 5.

38. Используя данные задачи 11 главы 5, определите односторонний 95% доверительный интервал для веса шоколадного батончика после вмешательства в процесс производства, указав, что средний вес в генеральной совокупности *не превышает* определенное значение.

39. Из списка 729 человек, участвовавших в круизе, для опроса случайно отобрали 130. Из них 112 человек ответили, что они очень довольны обслуживанием во время круиза. Найдите 95% доверительный интервал для процента тех людей в генеральной совокупности, которые довольны предоставленным обслуживанием.

40. Используя данные о качестве сельскохозяйственной продукции из задачи 29 главы 8.

    а) Найдите 95% доверительный интервал для среднего качества продуктов во всей генеральной совокупности.

    б) Найдите 95% интервал предсказания для результатов измерения качества продукции в следующем наблюдении.

    в) Найдите 99% доверительный интервал для среднего качества продукции во всей генеральной совокупности.

    г) Найдите 99% интервал предсказания для результатов измерения качества продукции в следующем наблюдении.

41. Ниже приведено содержание кофеина (в миллиграммах) в случайно отобранных чашках кофе.

    112,8; 86,4; 45,9; 110,3; 100,3; 93,3; 101,9; 115,7; 92,5; 117,3; 105,6; 81,6.

    а) Найдите односторонний 99% доверительный интервал для среднего значения генеральной совокупности содержания кофеина в чашке кофе, который утверждает, что "не менее чем ...".

    б) Найдите односторонний 99% интервал предсказания для содержания кофеина в следующей чашке кофе, также утверждающего "не менее чем ...".

Упражнения с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении А.

1. Будем считать эту базу данных генеральной совокупностью. Рассмотрим следующую выборку пяти номеров служащих в этой базе данных: 24, 54, 17, 34 и 53.

    а) На основе данных этой выборки найдите среднее, стандартное отклонение и стандартную ошибку для годовой заработной платы.

    б) Найдите 95% доверительный интервал для средней заработной платы во всей генеральной совокупности.

    в) Начертите график (аналогичный рис. 9.1.5), содержащий среднее выборки и доверительный интервал.

2. А теперь рассмотрим всю генеральную совокупность размеров заработной платы, что в реальной жизни сделать невозможно.

    а) Найдите среднее и стандартное отклонение генеральной совокупности и сравните их с выборочными оценками.

    б) Начертите для данной ситуации график (аналогичный рис. 9.1.1). Используйте $\sigma_{\bar{x}}$ как стандартное отклонение выборочного распределения.

    в) Попадает ли в этом случае среднее генеральной совокупности в доверительный интервал (из задачи 1)? Будет ли это среднее *всегда* находиться в этом интервале для всех случайных выборок? Почему да или почему нет?

3. Повторите упражнение 1, пп. "б" и "в", для 99% доверительного интервала. Попадает ли среднее генеральной совокупности годовой зарплаты в этот интервал?

4. Повторите упражнение 1, пп. "б" и "в", для 90% доверительного интервала. Попадает ли среднее генеральной совокупности годовой зарплаты в этот интервал?

5. Повторите упражнение 1, используя 95% доверительный интервал для следующей случайной выборки: номера служащих 4, 47, 45, 12 и 69. Дайте ответы на следующие вопросы:

    а) В реальной жизни что бы вы могли предпринять (если это возможно) в отношении того факта, что среднее генеральной совокупности не принадлежит доверительному интервалу?

    б) Постройте 99% и 99,9% доверительные интервалы. При каком уровне доверительности (если таковой существует) доверительный интервал будет достаточно большим, чтобы включать среднее генеральной совокупности?

6. Рассмотрим следующую случайную выборку из 15 номеров служащих в этой базе данных: 66, 37, 56, 11, 32, 23, 53, 43, 55, 25,7, 26, 36, 22 и 20.

    а) Найдите процент женщин в этой выборке.

    б) Найдите стандартную ошибку для процента женщин и поясните полученное значение.

    в) Какие у вас могли бы быть сомнения в использовании этой выборки и методов, изложенных в этой главе, при вычислении доверительного интервала для процента женщин?

7. Считая базу данных из приложения А случайной выборкой из более крупной генеральной совокупности, рассмотрим размер годовой зарплаты служащего.

    а) Найдите 95% доверительный интервал.

    б) Найдите 99% доверительный интервал.

8. Считая базу данных из приложения A случайной выборкой из более крупной генеральной совокупности, рассмотрим возраст служащего.

   а) Найдите 95% доверительный интервал.

   б) Найдите 90% доверительный интервал.

9. Считая базу данных из приложения A случайной выборкой из более крупной генеральной совокупности, рассмотрим стаж работы служащего.

   а) Найдите 95% доверительный интервал.

   б) Найдите 99,9% доверительный интервал.

10. Считая базу данных из приложения A случайной выборкой из более крупной генеральной совокупности, рассмотрим процент женщин. Постройте 95% доверительный интервал.

11. Считая базу данных из приложения A случайной выборкой из более крупной генеральной совокупности, рассмотрим процент служащих высокой квалификации (имеют уровень подготовки B или C). Постройте 99% доверительный интервал.

12. Рассматривая базу данных из приложения A как случайную выборку из более крупной генеральной совокупности

    а) Найдите односторонний 95% доверительный интервал для средней годовой зарплаты в генеральной совокупности, который утверждает, что зарплата по крайней мере составляет некоторую сумму.

    б) Найдите 99% односторонний доверительный интервал для п. "а".

    в) Найдите 95% односторонний доверительный интервал для среднего значения стажа работы служащего в генеральной совокупности, который утверждает, что стаж составляет *по крайней мере* определенное количество лет.

    г) Найдите 99% односторонний доверительный интервал для п. "в".

13. Рассматривая базу данных из приложения A как случайную выборку из идеализированной совокупности потенциальных сотрудников, которые могут быть приняты на работу

    а) Определите 95% интервал предсказания для стажа работы сотрудника, который будет нанят следующим. Почему этот интервал намного шире, чем доверительный интервал для среднего значения стажа во всей генеральной совокупности?

    б) Определите 95% интервал предсказания для возраста сотрудника, который будет нанят следующим.

Проекты

1. Получите оценку и стандартную ошибку этой оценки для *двух* показателей, важных для ваших интересов в бизнесе (используйте реальные данные или исходите из некоторых предположений). Для каждого из этих двух показателей постройте доверительный интервал и напишите небольшое пояснение к нему. В случае использования уровня доверительности, отличного от 95%, или в случае использования одностороннего доверительного интервала поясните, почему вы поступили именно так.

2. Получите оценку и стандартное отклонение этой оценки для *двух* показателей, важных для ваших интересов в бизнесе (используйте реальные данные или исходите из некоторых предположений). Для каждого из этих двух показателей постройте интервал предсказания и напишите небольшое пояснение к нему. В случае использования уровня предсказания, отличного от 95%, или в случае использования одностороннего интервала предсказания поясните, почему вы поступили именно так.

3. Найдите в Internet или в газете статью с результатами опроса общественного мнения. Напишите один абзац текста, посвященный одному из результатов опроса. Обязательно укажите объем выборки, процент и стандартную ошибку. Постройте сами двусторонний 95% доверительный интервал. Сравните ваши результаты с предельным значением ошибки опроса, если это значение указано в использованной вами статье.



Ситуация для анализа

Многообещающие результаты опроса относительно заказов фирменных товаров по каталогу

Недавно были подведены предварительные результаты исследования, касающегося проекта продаж фирменных товаров по каталогу, которые выглядят очень многообещающе! Средний планируемый размер заказа составил $53,94, т.е. на $15 больше ожидаемого. Руководитель группы от радости, вероятно, должен быть на седьмом небе: так как $53,94 за каждый заказ, полученный из 1 300 000 потенциально возможных адресов, приводит к продажам со средней суммарной цифрой более 70 миллионов долларов!

При подготовке совещания в ваши обязанности входит оценить, насколько корректно было проведено данное исследование. Исходная докладная записка, кроме суммы в $53,94, содержала мало деталей. Сделав несколько телефонных звонков, вы узнали, кто из служащих проделал основной объем работы в этом исследовании. Ниже приводится полученная вами информация. Случайная выборка была извлечена из соответствующей базы данных, содержащей адреса 600 000 обеспеченных людей, покупающих предметы роскоши по почте. По почте были разосланы 600 каталогов вместе с вопросниками. Также вы узнали, что 74 из этих 600 вопросников вернулись. В 9 из них было отмечено: "Да, я согласен до конца года заказать данные изделия на общую сумму _____ долларов". Указаны были следующие суммы: $7,97; $12,05; $29,27; $228,26; $2,28; $7,25; $114,39; $31,64 и $52,39.

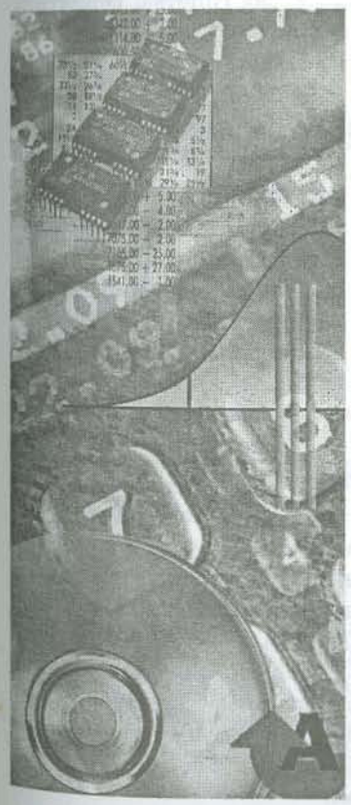
Теперь вам известно, что в размерах заказа существует существенный разброс. 95% доверительный интервал для среднего начинается с $3,10 и заканчивается $104,79. Умножив эти числа на количество возможных покупателей (1 300 000), получаем в качестве границ суммы от 4 030 000 до 136 227 000 долларов. Поэтому, даже приняв во внимание случайность, можно реально надеяться получить такие объемы продаж. А может быть, нет?

Вопросы для обсуждения

1. Можно ли было умножать среднее значение суммы заказа ($53,94) на количество (1 300 000) возможных покупателей?
2. Может, лучше умножить (как и предложено) граничные значения доверительного интервала на количество возможных покупателей?
3. А может, лучше умножать на количество адресов в основе выборки, которая использовалась для извлечения случайной выборки?
4. Может быть, вас еще что-либо смущает в этой ситуации?
5. Какой будет ваша наилучшая оценка с доверительными границами для потенциальных продаж по каталогу?

Проверка статистических гипотез: выбор между реальностью и совпадением

О нет! На линии настойчивый коммерческий агент пытается продать вам чудесную, повышающую выход готовой продукции добавку для увеличения производительности вашего нефтеперерабатывающего завода. Это кажется выгодной сделкой, но вы колеблетесь. Вы испытывали ее в течение недели и убедились в эффективности. Но это испытание не в полном объеме, а так как процесс может изменяться, то трудно сказать, получится из этого что-либо существенное или нет. Вам необходима объективная оценка, но вы знаете, что по телефону вы услышите только реплику: "Объем продукции возрос, не так ли? Ну, что я вам говорил? Если вы подпишете сегодня, мы дадим вам "зеленую улицу". Итак, вы подаете тайный сигнал своему секретарю, чтобы она ответила, что вы сейчас на собрании и позвоните позже.



Что же вас беспокоит? Да, объем продукции повышается. Но если даже специально не предпринимать никаких мер, то объем продукции также отклоняется каждый день, каждую неделю от вычисленного за долгий период работы среднего значения. Следовательно, объем продукции увеличивается по одной из двух причин: либо добавка действительно работает, либо это просто совпадение. Иными словами, независимо от наличия добавки, шансы получить такой недельный объем продукции, который превышает среднее значение за длительный период, составляют приблизительной 50 на 50, и в то же время шансы получить недельный объем продукции ниже этого среднего значения и также составляют приблизительно 50 на 50.

Теперь посмотрим на эту проблему с точки зрения коммерческого агента. Предположим, что эта добавка действительно бесполезна и никак не влияет на результат. Убеждаем менеджеров в 100 различных компаниях испытать эту добавку в течение недели. Примерно 50 менеджеров обнаружат, что объем продукции уменьшился, и с ними можно дальше не работать. Но другие 50 менеджеров обнаружат, что объем продукции слегка повысился. Может быть, некоторые из них даже заплатят большие деньги, чтобы продолжать использовать этот бесполезный продукт.

Вам необходимо проанализировать собранную информацию, чтобы определить, является ли повышение объема продукции на прошлой неделе *простым совпадением* (один вариант), или вы имеете убедительное доказательство того, что эта добавка действительно работает (другой вариант). Проверка статистических гипотез помогает решать подобные задачи.

Проверка статистических гипотез позволяет на основе имеющейся информации сделать выбор между двумя предположениями (которые называют *гипотезами*).[1] Эта процедура дает ответ на вопрос, являются ли наблюдаемые результаты простым совпадением (и их причиной можно с полным основанием считать случай), или они реальны. О проверке статистических гипотез иногда говорят как о способе использования статистики для принятия решений. Если смотреть на проблему шире, то можно рассматривать проверку гипотез как *один из компонентов* процесса принятия решений. Сама по себе проверка гипотез, вероятно, не может быть использована для принятия решения о покупке продукта, но она дает важную информацию об эффективности продукта.

10.1. Не все гипотезы одинаковы!

Гипотеза представляет собой некоторое утверждение об окружающем мире. Это утверждение относится к *генеральной совокупности*. Гипотеза не обязательно должна быть верной; она может быть либо верной, либо неверной, и для решения этого вопроса используют выборочные данные. Если все известно, то необходимости в проверке статистических гипотез нет. При наличии неопределенности проверка статистических гипотез помогает максимально использовать имеющуюся информацию.

Обычно мы будем рассматривать *две* гипотезы. Используя имеющиеся данные, мы будем решать, какая из этих двух гипотез более предпочтительна. Но эти гипотезы не взаимозаменяемы; каждая из них играет свою особую роль.

Нулевая гипотеза

Нулевая гипотеза, обозначается H_0 и представляет собой такое утверждение, которое принимается тогда, когда *нет убедительных аргументов для его отклонения*. Это очень выгодная позиция. Если ваши данные неполны или слишком неоднородны, то вы примете нулевую гипотезу, поскольку она имеет "презумпцию справедливости". Фактически нулевую гипотезу можно принять без реальных доказательств, поставив себя тем самым в довольно уязвимую позицию.

[1] Единственное число — *hypothesis*, множественное — *hypotheses*.

В значительной мере исходя из этого вы и определяете то, какую из двух гипотез вы будете считать нулевой.

Из двух гипотез нулевая часто является *более определенной*. Например, нулевая гипотеза может утверждать, что среднее совокупности точно равно некоторому определенному значению или что наблюдаемое различие имеет случайную природу. Чтобы было понятно, что гипотеза о случайности действительно является более определенной, заметим, что *неслучайности* могут характеризоваться разными структурами, а случайность подразумевает отсутствие структуры.

Исследовательская гипотеза

Исследовательскую гипотезу обозначают H_1 и принимают только тогда, когда есть убедительное статистическое доказательство, которое отвергает приемлемость нулевой гипотезы. Исследовательскую гипотезу также называют альтернативной гипотезой. Принятие альтернативной гипотезы представляет более сильную позицию, чем принятие нулевой гипотезы, так как она требует убедительного доказательства.

Часто исследовательская гипотеза представляет собой как бы скрытые планы исследователя, и нулевая гипотеза выдвигается только для того, чтобы быть отвергнутой. Окончательный результат тогда представляется так, что "это не случайность, и я могу объяснить это таким образом...". Это принятый способ проведения исследования. Поскольку многие люди имеют склонность к фантазиям, научное сообщество требует, чтобы перед публикацией полученных результатов была отвергнута нулевая гипотеза об их чистой случайности, что позволяет эффективно отсеивать многие абсолютно дикие идеи, не имеющие под собой никакого фактического основания. Это не является *гарантией* правильности всех полученных результатов, но это отсеивает многие некорректные идеи.

Определяя, какая из двух гипотез будет альтернативной, следует спросить себя: "Какая из гипотез *требует доказательств*?" Иначе говоря, необходимо определить, для принятия какой из гипотез требуются более убедительные доказательства. Эта гипотеза и должна стать альтернативной исследовательской гипотезой. Не пренебрегайте личными интересами! Не стесняйтесь переложить бремя доказательств на тех, кто пытается продать вам свою продукцию. Пусть они доказывают свои утверждения!

О чем свидетельствует результат

Есть два возможных результата проверки гипотезы. Для удобства мы представим их следующим образом.

Результаты проверки гипотезы		
Или	Принять нулевую гипотезу H_0 в качестве приемлемой возможности	Слабое заключение; отсутствует значимый результат
Или	Отвергнуть нулевую гипотезу H_0 и принять альтернативную гипотезу H_1	Сильное заключение; есть значимый результат

Заметим, что *никогда* не говорят об отклонении альтернативной гипотезы. Это обусловлено тем, что нулевая гипотеза имеет предпочтительный статус быть принятой без доказательств. Принятие нулевой гипотезы просто означает, что нет достаточных доказательств для ее опровержения. Мы решаем "принять" нулевую гипотезу не будучи обязательно уверенными в ее истинности. Принимая ее как возможный сценарий, который мог бы привести к получению таких данных, мы, тем не менее, признаем, что существует много других в такой же мере возможных сценариев, которые *близки к* нулевой гипотезе и которые также могли привести к получению таких данных. Например, когда мы принимаем нулевую гипотезу о том, что среднее совокупности равно $2000, то обычно не исключаем того, что среднее может быть равно $2001 или $1999. По этой причине некоторые статистики предпочитают говорить, что у нас "недостаточно данных, чтобы отклонить" нулевую гипотезу, а не просто, что мы "принимаем" ее.

Можно рассматривать нулевую гипотезу в терминах уголовного права. Нулевая гипотеза утверждает "невиновен", в то время как альтернативная гипотеза утверждает "виновен". Поскольку наша правовая система основана на принципе "невиновен, пока не доказана виновность", есть смысл обозначить гипотезы именно так. Принятие нулевой гипотезы о невиновности означает, что нет достаточных доказательств для осуждения, но в то же время это не является действительным доказательством невиновности. В то же время отклонение нулевой гипотезы и принятие альтернативной гипотезы о виновности говорит о том, что существует достаточно доказательств, которые исключают возможность невиновности и убеждают в виновности. Мы не должны показывать отсутствие вины для доказательства невиновности, но мы должны показать отсутствие невиновности для доказательства виновности.

Примеры гипотез

Ниже приведены примеры нулевых и альтернативных гипотез, сформулированных относительно генеральной совокупности. Обратите внимание, что в каждом случае обе гипотезы не могут быть верными одновременно и чтобы выбрать, какую из них следует принять, необходимо использовать данные.

1. *Ситуация.* Случайно отобранная группа из 200 человек посмотрела рекламу; после этого регистрируется количество людей из этой группы, которые в течение следующей недели купили рекламируемый продукт.

   Нулевая гипотеза. Реклама не имела никакого эффекта. Другими словами, процент покупателей среди тех в генеральной совокупности, кто видел рекламу, *в точности равен* проценту покупателей среди тех в генеральной совокупности, кто *не* видел ее (т.е. равен обычному уровню продаж). Из прошлого опыта известно, что этот обычный уровень продаж составляет 19,3%.

   Альтернативная гипотеза. Реклама имеет эффект. Это значит, что процент покупателей среди тех в генеральной совокупности, кто видел рекламное объявление, *отличается* от обычного уровня продаж, равного 19,3%, и представляющего тех покупателей в генеральной совокупности, кто *не* видел рекламу.

   Обсуждение. Обратите внимание, что эти гипотезы представляют собой предположения относительно *генеральной совокупности* в целом, а не относительно выборки из 200 человек. Выборочные данные, собранные в результате наблюдения поведения 200 случайно отобранных человек, помогут решить, какую из гипотез принять. Поскольку нулевая гипотеза содержит точное значение процента, она является более определенной, чем альтернативная гипотеза, которая содержит утверждение о более широком диапазоне (т.е. о любом значении, отличном от 19,3%). Кроме того, заметим, что если вы примете решение, что реклама была эффективной, вы сделаете более строгое утверждение, так как это альтернативная гипотеза. Это так, как если бы вы заявили: "Прекрасно. Если эта реклама работает так хорошо, как мы думаем, давайте это докажем. Или, с другой стороны, если эта реклама будет иметь катастрофические последствия для продаж, давайте это также докажем".

2. *Ситуация.* Вы оцениваете добавку, улучшающую объем выпуска продукции, описанную в начале этой главы.

   Нулевая гипотеза. Добавка в долгосрочном плане не оказывает влияния на объем выпускаемой продукции, величина которого известна из прошлого опыта.

   Альтернативная гипотеза. Добавка оказывает некоторое долгосрочное влияние на объем выпускаемой продукции.

   Обсуждение. Нулевая гипотеза является более определенной. Обе гипотезы формулируются относительно генеральной совокупности (объем продукции за длительный период времени), а не только относительно результатов работы за прошедшую неделю (выборка). Вы должны опровергнуть, что добавка неэффективна, и убедиться в противоположном, что потребует дополнительных доказательств. Производители добавки должны представить доказательства и продемонстрировать ее эффективность. Это не ваша забота доказывать им, что добавка *не* является эффективной.

3. *Ситуация.* Вашей фирме предъявлен иск в дискриминации сотрудников по полу, и вы анализируете документы, представленные другой стороной. Они включают проверку статистической гипотезы, основанную на данных о заработной плате мужчин и женщин, которая утверждает наличие "очень значительной разницы" между размерами средней заработной платы мужчин и у женщин.

   Нулевая гипотеза. Размеры заработной платы мужчин и женщин равны, если не принимать во внимание случайные отклонения. Иными словами, реальная разница в размерах заработной платы мужчин и женщин не намного бы изменилась, если бы мы сложили все заработные платы в шляпу, хорошо их перемешали и раздали сотрудникам без учета их пола.

   Альтернативная гипотеза. Различие в размерах заработной платы мужчин и женщин превышает простую случайность.

   Обсуждение. Обратите внимание, что здесь используют идеализированную совокупность. Ввиду того, что эти служащие никак не могут рассматри-

ваться как случайная выборка, гипотеза относится к некоторой идеализированной совокупности, которая представляет собой мужчин и женщин, равных с точки зрения размера заработной платы и таких, что все имеющие место различия в размерах заработной платы могут быть объяснены случайностью распределения заработной платы между отдельными людьми. Если будет отвергнута нулевая гипотеза и принята альтернативная, то ваша фирма будет иметь проблемы. Такое сильное заключение будет работать против вас. Но не все потеряно. Не забывайте, что статистические методы говорят в основном только о числах, а не о том, *почему* эти числа именно такие. Разница в заработной плате может быть обусловлена непосредственно дискриминацией по полу, а может, и другими факторами, такими как образование, опыт и способности. Проверка статистической гипотезы, рассматривающей только пол и размер заработной платы, не может показать, какие именно факторы повлияли на эту разницу.[2] Кроме того, результаты проверки гипотезы могут быть неверными, так как использование статистических методов всегда связано с наличием ошибок.

10.2. Проверка гипотезы о равенстве среднего генеральной совокупности некоторому заданному значению

Один из самых простых случаев проверки статистической гипотезы заключается в проверке равенства между средним генеральной совокупности и некоторым заданным значением. Заданное значение представляет собой некоторое фиксированное число μ_0, полученное не из выборочных данных. Гипотезы имеют следующий вид.

$$H_0 : \mu = \mu_0$$

Нулевая гипотеза утверждает, что неизвестное среднее значение генеральной совокупности μ *в точности равно* заданному значению μ_0.

$$H_1 : \mu \neq \mu_0$$

Альтернативная гипотеза утверждает, что неизвестное среднее значение генеральной совокупности μ *не равно* заданному значению μ_0.

Это двусторонняя проверка, поскольку альтернативная гипотеза рассматривает возможность того, что значение среднего генеральной совокупности может быть расположено по обе стороны (как больше, так и меньше) от заданного значения μ_0.[3] Обратите внимание, что фактически здесь фигурируют *три* различных числа, имеющих отношение к среднему:

[2] В дальнейшем вы узнаете, как *множественная регрессия* позволяет учесть влияние других факторов (таких как образование и стаж работы) и дать *скорректированную оценку* влияния пола на заработную плату при условии, что другие факторы являются постоянными.

[3] Об *односторонней* проверке статистических гипотез вы узнаете из раздела 10.4.

- μ — неизвестное среднее генеральной совокупности, которое вас интересует;
- μ_0 — заданное значение, в отношении которого проверяют гипотезу;
- $\bar{X}$ — известное выборочное среднее, которое используют для вынесения решения о принятии гипотезы. Из указанных трех чисел только это значение является случайной величиной, так как оно рассчитано из данных выборки. Заметим, что $\bar{X}$ является оценкой и, следовательно, представляет μ.

Проверка гипотезы заключается в сравнении двух известных величин $\bar{X}$ и μ_0. Если эти значения отличаются сильнее, чем можно было бы ожидать исходя из случайности, то нулевую гипотезу $\mu = \mu_0$ отклоняют, так как $\bar{X}$ предоставляет информацию о неизвестном среднем μ. Если значения $\bar{X}$ и μ_0 достаточно близки, то нулевую гипотезу $\mu = \mu_0$ принимают. Но что означает "значения близки"? Где находится необходимая граница? Близость должна определяться на основе значения $S_{\bar{X}}$, поскольку эта стандартная ошибка определяет степень случайности $\bar{X}$. Таким образом, если $\bar{X}$ и μ_0 отстоят друг от друга на расстоянии достаточного количества стандартных ошибок, то это является убедительным доказательством того, что μ не равно μ_0.

Существуют два различных метода проверки гипотезы и получения результата. Первый метод использует доверительные интервалы, о которых шла речь в предыдущей главе. Это более простой метод, потому что (а) вы уже знаете, как строить и интерпретировать доверительный интервал, и (б) доверительный интервал интерпретируется непосредственно, поскольку он выражен в тех же единицах измерения, что и данные (например, в долларах, количестве людей, количестве поломок). Второй метод (основанный на t-статистике) является более традиционным, но интуитивно менее понятным, поскольку заключается в том, чтобы вычислить показатель, измеренный не в тех же единицах, что и данные, сравнить полученное значение с соответствующим критическим значением из t-таблицы и затем сделать вывод.

Не имеет значения, какой метод использовать для проверки гипотезы (на основе доверительного интервала или на основе t-статистики), поскольку оба метода дают всегда одинаковые результаты. Может, вам захочется чаще использовать метод на основе доверительного интервала, так как он быстрее, проще и дает больше информации о ситуации. В то же время вы можете захотеть узнать об использовании метода на основе t-статистики, так как именно этот метод обычно применяют на практике. Поскольку два этих метода приводят к одному результату, оба их называют *t-тестом*.

Использование доверительных интервалов: простой способ

Рассмотрим проверку нулевой гипотезы $H_0 : \mu = \mu_0$ против альтернативной гипотезы $H_0 : \mu \neq \mu_0$ на основе данных случайной выборки из генеральной совокупности. Сначала обычным образом (см. главу 9) исходя из значений $\bar{X}$ и $S_{\bar{X}}$ строим 95% доверительный интервал. Затем смотрим, попадает ли значение μ_0 в этот интервал. Если значение μ_0 находится за пределами доверительного интервала, то μ_0 не может рассматриваться как допустимое значение среднего генеральной совокупности, а значит, следует принять альтернативную гипотезу; в противном случае принимается нулевая гипотеза. Этот подход проиллюстрирован на рис. 10.2.1.

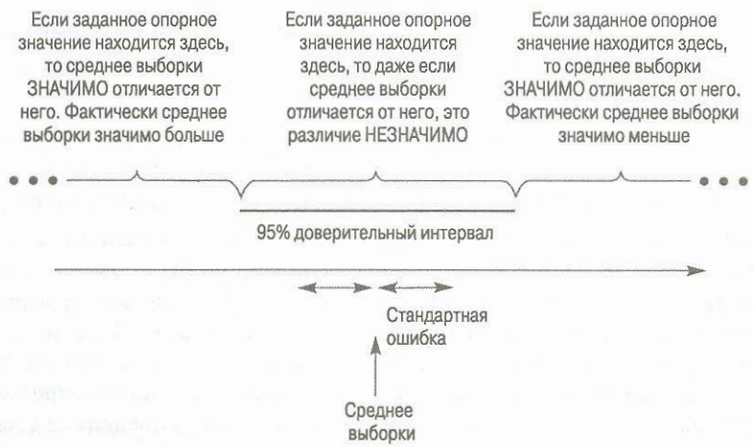


Рис. 10.2.1. Проверка гипотезы о среднем генеральной совокупности, основанная на доверительном интервале. Вопрос состоит в том, можно ли принять утверждение о равенстве среднего генеральной совокупности и заданного значения. Если это заданное значение попадает в доверительный интервал, тогда это утверждение приемлемо. Если заданное значение находится вне интервала, то принимается решение о том, что это значение не может быть средним генеральной совокупности

Есть несколько эквивалентных способов описать результат такой проверки статистической гипотезы. В каждом случае принятое решение может быть сформулировано так, как показано в табл. 10.2.1.

Таблица 10.2.1. Принятие решения относительно проверки гипотезы о среднем генеральной совокупности

Если заданное значение μ_0 находится в доверительном интервале от $\bar{X} - tS_{\bar{X}}$ до $\bar{X} + tS_{\bar{X}}$, то:
Принять нулевую гипотезу H_0 как допустимую возможность.
Не принимать альтернативную гипотезу H_1.
Выборочное среднее $\bar{X}$ незначимо отличается от заданного значения μ_0.
Наблюдаемая разница между выборочным средним $\bar{X}$ и заданным значением μ_0 может быть обусловлена чистой случайностью.
Результат проверки не является статистически значимым.
(Все перечисленные выше утверждения эквивалентны.)
Если заданное значение μ_0 не находится в доверительном интервале от $\bar{X} - tS_{\bar{X}}$ до $\bar{X} + tS_{\bar{X}}$, то:
Принять альтернативную гипотезу H_1.
Отклонить нулевую гипотезу H_0.
Выборочное среднее $\bar{X}$ значимо отличается от заданного значения μ_0.
Наблюдаемая разница между выборочным средним $\bar{X}$ и заданным значением μ_0 не может быть обусловлена только лишь случайностью.
Результат проверки является статистически значимым.
(Все перечисленные выше утверждения эквивалентны.)

Почему этот метод работает? Вспомним, что в соответствии с формулировкой доверительного интервала вероятность того, что μ находится в (случайном) доверительном интервале, равна 0,95. Допустим на мгновение, что нулевая гипотеза верна и $\mu = \mu_0$. Тогда вероятность того, что μ_0 находится в доверительном интервале, также равна 0,95. Это говорит о том, что если нулевая гипотеза верна, то принятое решение будет корректным (приблизительно) в 95% случаев и будет неверным только приблизительно в 5% случаев. В этом смысле мы имеем процесс принятия решений с точной, контролируемой вероятностью. В разделе 10.3 мы подробно рассмотрим различные типы ошибок, возникающих при проверке статистических гипотез.

Пример. Действительно ли добавка "увеличивает объем продукции"?

Вспомним добавку, увеличивающую (предположительно) объем продукции, покупку которой мы рассматривали в начале этой главы. Будем считать, что в табл. 10.2.2 приведены основные данные, характеризующие эту задачу.

Данные содержат 7 наблюдений объема продукции при условии использования добавки. Генеральная совокупность состоит из всех возможных дневных объемов продукции, полученных с использованием добавки; в частности, среднее генеральной совокупности μ представляет среднее значение объема продукции, полученной за длительный период в условиях применения добавки (это значение неизвестно и поэтому в таблице не приведено). Выборочное среднее $\bar{X}$ представляет наилучшую оценку μ.

Действительно, данные в таблице выглядят так, как будто добавка эффективна. Средний дневной объем продукции, достигнутый с применением добавки ($\bar{X} = 39{,}6$ тонны), на 7,5 тонны выше ожидаемого среднего дневного объема продукции в отсутствие добавки, рассчитанного за предыдущий длительный период ($\mu_0 = 32{,}1$ тонны). Это не удивительно. При проверке гипотезы заданное значение почти никогда точно не равно наблюдаемому значению (в нашем случае $\bar{X}$). Вопрос заключается в том, обусловлена ли эта разница только случайностью. Гистограмма данных с обозначенными на ней выборочным средним и заданным значениями показана на рис. 10.2.2.

При подготовке к проверке статистических гипотез выдвигают гипотезы, которые теперь можно сформулировать непосредственно с использованием заданного значения $\mu_0 = 32{,}1$ тонны. (Теперь больше нет необходимости использовать в формальной записи гипотез символическое μ_0 вместо его известного значения.) Гипотезы формулируются следующим образом.

$$H_0 : \mu = 32{,}1 \text{ тонны}$$

Нулевая гипотеза утверждает, что при использовании добавки неизвестное среднее значение объема продукции за длительный период времени μ точно *равно* заданному значению $\mu_0 = 32{,}1$ тонны (объему продукции без использования добавки).

$$H_1 : \mu \neq 32{,}1 \text{ тонны}$$

Альтернативная гипотеза утверждает, что при использовании добавки неизвестное среднее значение объема продукции за длительный период времени μ *не равно* заданному значению $\mu_0 = 32{,}1$ тонны (объему продукции без использования добавки).

Далее, для проверки гипотезы вычислим обычным способом доверительный интервал, используя значение $t = 2{,}447$ из t-таблицы для $n - 1 = 6$ степеней свободы.

Мы на 95% уверены, что при использовании добавки среднее значение объема продукции за длительный период времени находится между 29,3 и 49,9 тонны.

Таблица 10.2.2. Основные данные, характеризующие добавку, "увеличивающую объем продукции"

Средний ежедневный объем продукции на прошлой неделе	$\bar{X}$	39,6 тонны
Стандартная ошибка	$S_{\bar{x}}$	4,2 тонны
Размер выборки	n	7 дней
Среднее значение ежедневного объема продукции за длительный период (без использования добавки)	μ_0	32,1 тонны

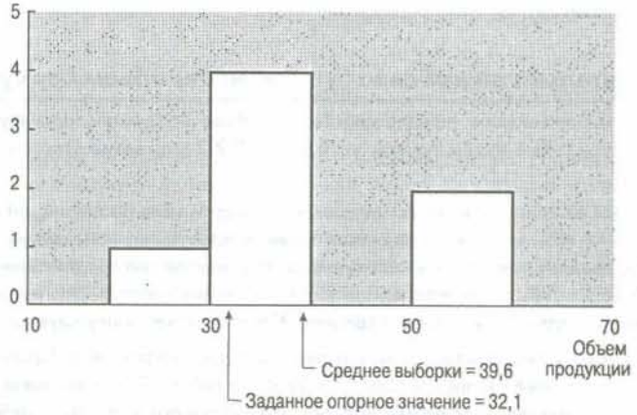


Рис. 10.2.2. Гистограмма объемов продукции за 7 дней работы с применением добавки. Среднее выборки, обобщающее эти данные, выше заданного значения. Но является ли эта разница значимой? Проверка гипотезы покажет, действительно ли эта гистограмма выборки, которая может быть извлечена из генеральной совокупности со средним, равным заданному значению

Наконец, чтобы действительно осуществить процедуру проверки гипотезы, следует просто проверить, находится ли заданное значение $\mu_0 = 32{,}1$ в пределах доверительного интервала или нет.[4] Это значение находится в интервале, поскольку число 32,1 лежит между числами 29,3 и 49,9. Иными словами, утверждение $29{,}3 \le 32{,}1 \le 49{,}9$ является справедливым. Результат проверки гипотезы приведен в табл. 10.2.3.

Среднее значение дневного объема продукции при использовании добавки, равное $\bar{X} = 39{,}6$ тонны, несущественно отличается от среднего значения объема продукции за длительный период времени без применения добавки, которое равно $\mu_0 = 32{,}1$ тонны. Этот результат неубедительный и неоднозначный. Вы не имеете четкого доказательства в пользу добавки. Когда вы в следующий раз будете разговаривать с настойчивым коммерческим агентом, пытающимся продать вам добавку, то можете с уверенностью сказать ему, что *такое повышение объема продукции не является значимым* и вы не убеждены в эффективности добавки.

Доказала ли проверка гипотезы неэффективность добавки? Нет. Добавка может быть и эффективной. У вас нет убедительных доказательств ни ее эффективности, ни ее неэффективности.

[4] Не нужно проверять, принадлежит ли интервалу значение $\bar{X}$, поскольку оно всегда находится в этом интервале. Вопрос заключается в том, принадлежит ли интервалу *известное значение* μ_0.

Что еще следует предпринять для решения этой проблемы? Ваш коммерческий агент мог бы предложить использовать эту добавку еще в течение месяца — конечно, бесплатно, — чтобы посмотреть, будет ли достаточно убедительной дополнительная информация. Либо, если у вас хватит нервов и решимости, вы можете предложить провести такую проверку ему.

Таблица 10.2.3. Результат проверки гипотезы о добавке, "повышающей объем продукции"

Поскольку заданное значение $\mu = 32{,}1$ тонны находится в пределах доверительного интервала от 29,3 до 49,9 тонны, то:

Принимаем нулевую гипотезу $H_0 : \mu = 32{,}1$ как допустимую возможность.

Не принимаем альтернативную гипотезу $H_1 : \mu \neq 32{,}1$ тонны.

Выборочное среднее значение объема продукции $\bar{X} = 39{,}6$ незначимо отличается от заданного значения $\mu_0 = 32{,}1$.

Наблюдаемая разница между выборочным средним значением объема продукции $\bar{X} = 39{,}6$ и заданным значением $\mu_0 = 32{,}1$ может быть обусловлена только лишь чистой случайностью.

Результат проверки не является статистически значимым.

(Все предшествующие утверждения эквивалентны.)

Пример. *Передача акций наемным работникам и качество продукции*

Существует много способов достичь высокого качества продукции. Один из них заключается в предоставлении работникам возможности получать часть прибыли в зависимости от результатов их работы. Согласно этой теории, служащие, которые имеют часть акций компании, имеют прямое вознаграждение (повышенный размер дивидендов и повышенную стоимость акций) за свою высококачественную продукцию. Это является дополнительным стимулом для служащих самостоятельно повышать качество своей работы.

Было проведено исследование, чтобы определить, как менеджеры 343 фирм смотрят на эффективность различных видов программ трудовых отношений служащих.[5] Обратите внимание, что это исследование измеряет не эффективность этих программ, а восприятие этой эффективности менеджерами. Тем не менее, учитывая, что многие менеджеры имеют непосредственный опыт работы с такими программами, результаты исследования являются интересными и полезными.

Каждый из 343 менеджеров оценил влияние передачи акций на качество продукции, используя шкалу оценки от –2 до 2, где –2 означало "сильное отрицательное влияние", а 2 — "сильное положительное влияние". Среднее значение оценки составило $\bar{X} = 0{,}35$ со стандартной ошибкой $S_{\bar{x}} = 0{,}14$.

Восприятие влияния передачи акций служащим на качество продукции компании (оценки в баллах от –2 до 2)

Средний балл	$\bar{X}$	0,35
Стандартная ошибка	$S_{\bar{x}}$	0,14
Размер выборки	n	343

Возникает вопрос: действительно ли менеджеры смотрят на передачу акций служащим компании как на эффективный метод повышения качества продукции? Ответ можно получить, выполнив проверку гипотезы. Почему нельзя просто использовать факт, что средний балл, равный 0,35, очевидно, свидетельствует о том, что менеджеры считают такой метод полезным. Да потому, что это результат для выборки менеджеров, и он может представлять, а может и не представлять генеральную совокупность менеджеров в це-

[5] Voos P. B., "Managerial Perceptions of the Economic Impact of Labour Relations Programs". *Industrial and Labour Relations Review* 40 (1987), p. 195–208.

лом. Чтобы сделать заключение о взглядах менеджеров в целом исходя из среднего для выборки менеджеров, следует выполнить проверку гипотезы.

Поскольку мы хотим убедиться в том, что передача акций служащим компании действительно дает результат, нам необходимы соответствующие доказательства и поэтому мы выдвинем именно такую альтернативную исследовательскую гипотезу. Нулевая гипотеза будет утверждать, что обладание акциями не влияет на качество продукции. Если мы определим μ как среднее значение оценок для большей генеральной совокупности менеджеров фирм, аналогичных фирмам в нашей выборке, то гипотезы можно сформулировать следующим образом.

$$H_0 : \mu = 0$$

Нулевая гипотеза утверждает, что неизвестное среднее значение оценок всех менеджеров μ точно равно заданному значению $\mu_0 = 0$.

$$H_1 : \mu \neq 0$$

Альтернативная гипотеза утверждает, что неизвестное среднее значение оценок всех менеджеров μ не равно заданному значению $\mu_0 = 0$.

Далее, для проверки гипотезы вычислим обычным способом доверительный интервал, используя значение t = 1,960 из t-таблицы.

Мы на 95% уверены, что среднее значение оценок всех менеджеров находится между 0,08 и 0,62.

Наконец, чтобы осуществить процедуру проверки гипотезы, следует просто проверить, находится ли заданное значение $\mu_0 = 0$ в пределах доверительного интервала или нет. Это значение не находится в интервале, потому что число 0 не лежит между числами 0,08 и 0,62. Результаты вашего t-теста приведены в табл. 10.2.4.

Таблица 10.2.4. Результаты проверки гипотезы относительно того, влияет ли передача акций служащим на качество продукции

Поскольку значение заданной опорной величины $\mu_0 = 0$ не находится в доверительном интервале от 0,08 до 0,62, то:
Принять альтернативную гипотезу $H_1 : \mu \neq 0$.
Отклонить нулевую гипотезу $H_1 : \mu = 0$.
Выборочное среднее значение оценки $\bar{X} = 0{,}35$ незначимо отличается от заданного значения $\mu_0 = 0$.
Наблюдаемая разница между выборочным средним значением оценок $\bar{X} = 0{,}35$ и заданным значением $\mu_0 = 0$ не может быть обусловлена только лишь случайностью.
Результат является статистически значимым.
(Все предыдущие утверждения эквивалентны.)

Исходя из мнения менеджеров влияние обладания служащими акциями на качество продукции имеет *статистически значимый позитивный эффект*.[6] Этот результат является доказанным. Вы имеете убедительное доказательство, что менеджеры в целом считают, что владение акциями служащими компании оказывает эффективное влияние на качество продукции. Даже несмотря на то, что полученная оценка относительно низка (0,35 при шкале от –2 до 2), менеджеры в целом согласны, что такой вариант является полезным.

[6] Можно утверждать, что *эффект позитивный*, так как (1) результат является статистически значимым, и (2) значение эффекта, выраженное через $\bar{X}$, является положительным числом (т.е. $\bar{X}$ — это положительное число, большее, чем $\mu_0 = 0$).

Действительно ли проверка доказала это абсолютно? Если опросить большее количество менеджеров, будет ли средняя оценка также положительной? Не обязательно. Абсолютное доказательство вообще невозможно тогда, когда есть даже очень незначительная случайность. У вас есть убедительное доказательство, но не абсолютное. Таким образом, приняв альтернативную гипотезу, мы могли допустить ошибку, хотя это маловероятно. Эти вездесущие ошибки рассматриваются в разделе 10.3.

В случае биномиального распределения проверяют, равен ли процент в генеральной совокупности π заданному значению π_0 при условии, что размер выборки n не слишком мал. Процедура проверки не сильно отличается от проверки для среднего совокупности, потому что, если имеется оценка и ее стандартная ошибка, доверительный интервал и t-статистику получают аналогичным способом. Общие черты, характерные для проверки гипотез при нормальном и биномиальном распределении, приведены в следующей таблице.

	Нормальное распределение	Биномиальное распределение
Среднее совокупности	μ	π
Заданное значение	μ_0	π_0
Нулевая гипотеза	$H_0 : \mu = \mu_0$	$H_0 : \pi = \pi_0$
Альтернативная гипотеза	$H_1 : \mu \neq \mu_0$	$H_1 : \pi \neq \pi_0$
Данные	$X_1, \ldots, X_n$	X наступлений события в n испытаниях
Оценка	$\bar{X}$	$p = X / n$
Стандартная ошибка	$S_{\bar{X}} = S\sqrt{n}$	$S_p = \sqrt{p(1-p)/n}$
Доверительный интервал	от $\bar{X} - tS_{\bar{X}}$ до $\bar{X} + tS_{\bar{X}}$	от $p - tS_p$ до $p + tS_p$
t-статистика	$t = (\bar{X} - \mu_0) / S_{\bar{X}}$	$t = (p - \pi_0) / S_p$

Пример. *Улучшение продукции (случай биномиального распределения)*

Одна из загадок производства электронных компьютерных микросхем состоит в том, что до их тестирования нельзя быть уверенным в хороших результатах. И здесь вы определяете, что качество одних микросхем недопустимо низкое, других — хорошее, а некоторые микросхемы имеют особенно высокое качество. Эти высококачественные микросхемы отделяют от остальных и продают как "особо быстрые", поскольку они обрабатывают информацию быстрее других.

Вам необходимо усовершенствовать производственный процесс до такой степени, чтобы более 10% продукции можно было продавать как "особо быстрые" микросхемы. Взяв за основу выборку из 500 недавно изготовленных микросхем, вы планируете выполнить проверку статистической гипотезы, чтобы посмотреть, достигнута ли поставленная цель, или до нее еще далеко, или цель уже совсем близко.

Поскольку недавно был модернизирован производственный процесс, вы надеетесь на хорошие результаты проверки. В выборке оказалось 58 особо быстрых микросхем, что в процентном отношении составило 11,6%, а это выше 10%. Но действительно ли процент необходимых вам микросхем значимо превысил заданное граничное значение или это просто случайная удача? Прежде чем праздновать победу, хотелось бы знать это точно.

Будем считать, что имеет место ситуация биномиального распределения, в которой каждая из микросхем может либо быть особо быстрой, либо не быть такой. Биномиальная вероятность π представляет для микросхемы вероятность быть особо быстрой. Размер выборки — $n = 500$, наблюдаемая частота — $X = 58$ и доля (выраженная в процентах) в выборке составляет $p = 11{,}6\%$. Заданное значение — $\pi_0 = 10\%$.

Проверка статистической гипотезы в случае биномиального распределения (с достаточно большим n) фактически не отличается от проверки гипотезы в случае количественных данных. В каждом из этих двух случаев имеем оценку ($\bar{X}$ или p), стандартную ошибку ($S_{\bar{x}}$ или Sp) и заданное значение (μ_0 или π_0). Гипотезы для биномиального случая формулируются следующим образом.

$$H_0 : \pi = 10\%$$

Нулевая гипотеза утверждает, что особо быстрые микросхемы составляют 10% от всего объема продукции.

$$H_1 : \pi \neq 10\%$$

Альтернативная гипотеза утверждает, что доля особо быстрых микросхем отличается от 10%: либо выше (ура! Время праздновать!), либо ниже (ой-ой, необходимо что-то менять).

Исходя из стандартной ошибки $S_p = \sqrt{p(1-p)/n} = 0{,}0143$ и табличного значения t = 1,960 вычисляем обычным для биномиального распределения способом 95% доверительный интервал. Получаем, что доверительный интервал находится в пределах от 8,8 до 14,4%.

Мы на 95% уверены, что доля особо быстрых микросхем в объеме всей продукции находится между 8,8 и 14,4%.

Теперь, чтобы проверить гипотезу, необходимо просто посмотреть, попадает заданное значение $\pi_0 = 10\%$ в интервал или нет. Это значение *попадает* в интервал, поскольку 10% лежит между 8,8 и 14,4%. Результат t-теста показан в табл. 10.2.5.

Наблюдаемое значение доли особо быстрых микросхем (в процентах) статистически незначимо отличается от 10%. У вас нет достаточно информации, чтобы на полном основании заявить, выше эта доля или ниже заданного значения. У вас нет окончательного вывода. Хотя величина 11,6% выглядит как хорошая доля (и фактически превышает заданное целевое значение, равное 10%), это значение доли незначимо отличается от целевого значения. Поскольку значение 11,6% может отличаться от 10% просто случайно, у вас нет строгого доказательства, что поставленная цель достигнута.

Помните, что вы занимаетесь статистическими выводами. Вас интересуют не только именно эти 500 микросхем. Хотелось бы знать о доле особо быстрых микросхем в продукции, выпущенной за длительный период на таком же оборудовании, на каком изготовлены эти микросхемы. Статистический вывод свидетельствует, что значение доли особо быстрых микросхем так близко к 10%, что невозможно сказать, достигнута ли уже поставленная цель.

Можно попытаться собрать дополнительные данные из завтрашнего выпуска продукции, чтобы посмотреть, позволяет ли эта дополнительная информация показать, что поставленная цель достигнута (принимая альтернативную гипотезу, вы утверждаете, что наблюдаемая доля *превышает* 10%). В то же время, вместо того, чтобы только собирать и анализировать данные, можно было бы подстраховаться, дополнительно модернизировав оборудование.

Таблица 10.2.5. Результаты проверки гипотезы о производстве микросхем

Поскольку заданное значение $\pi_0 = 10\%$ находится в пределах доверительного интервала от 8,8% до 14,4%, то:
Принять нулевую гипотезу $H_0 : \pi = 10\%$ как приемлемую возможность.
Не принимать альтернативную гипотезу $H_1 : \pi \neq 10\%$.
Выборочное значение доли $p = 11{,}6\%$ незначимо отличается от заданного значения $\pi_0 = 10\%$.
Наблюдаемая разница между выборочным значением доли $p = 11{,}6\%$ и заданным значением $\pi_0 = 10\%$ может быть обусловлена только лишь случайностью.
Результат проверки не является статистически значимым.
(Все предшествующие утверждения эквивалентны.)

t-статистика: способ другой, результат тот же

Другой способ двусторонней проверки гипотезы о среднем генеральной совокупности состоит в том, чтобы сначала вычислить t-статистику по формуле $(\bar{X} - \mu_0)/S_{\bar{x}}$, а затем, используя t таблицу, решить, какую из гипотез следует принять. Результат всегда будет таким же, как и при проверке методом доверительного интервала, поэтому неважно, какой из этих двух методов вы используете. Процедура проверки статистической гипотезы сравнения среднего генеральной совокупности с заданным значением исходя из значений $\bar{X}$ и $S_{\bar{x}}$ (использование обоих указанных методов) называется t-тестом Стьюдента, или просто t-тестом. (Используют также названия t-критерий Стьюдента и t-критерий. — *Прим. ред.*). Имя *Стьюдент* использовал В. С. Госсет, главный пивовар фирмы Guinness, при публикации первой статьи, в которой он вместо таблицы нормального распределения использовал t-таблицу (которую он первым и предложил), скорректированную для использования стандартного отклонения выборки S вместо неизвестного стандартного отклонения генеральной совокупности σ в условиях небольшого размера выборки n.[7]

В соответствии с общим походом проверка статистической гипотезы начинается с того, что на основе данных, содержащих наилучшую имеющуюся информацию, для установления различий между двумя гипотезами вычисляют величину, которую называют тест-статистикой. Далее эту тест-статистику (например, t-статистику) сравнивают с подходящим критическим значением, взятым из стандартной таблицы критических значений (например, t-таблицы), чтобы определить, какую гипотезу принять. В более сложных ситуациях, чем просто проверка гипотезы о среднем генеральной совокупности, могут потребоваться определенные творческие усилия, чтобы (1) подобрать тест-статистику, использующую информацию из выборки наиболее эффективно, и (2) найти подходящее критическое значение. При этом либо критическое значение определяют исходя из теоретических соображений (как в случае с t-таблицей), либо, как это все чаще делают в последнее время, специально вычисляют критические значения с помощью компьютеров для каждой отдельной ситуации.

Существуют две различные величины, которые обозначают буквой t. Критическое t-значение представляет собой число $t_{табл.}$, которое находят в t-таблице и которое никак не связано с выборочными данными. С другой стороны, t-статистика является тест-статистикой и показывает, сколько стандартных ошибок находится между μ_0 и $\bar{X}$.

t-статистика

Для распределения количественной переменной:
$$t_{статистика} = \frac{\bar{X} - \mu_0}{S_{\bar{x}}}.$$

Для биномиального распределения:
$$t_{статистика} = \frac{p - \pi_0}{S_p}.$$

[7] Student, "The Probable Error of a Mean", *Biometrika*, 6 (1908), p. 1–25.

Процедура t-теста использует обе эти величины, сравнивая t-статистику, вычисленную на данных, с t-значением, найденным по t-таблице. Результат проверки гипотезы сформулирован в табл. 10.2.6.

Абсолютное значение числа, которое записывают, размещая число между двумя вертикальными линиями, вычисляют путем удаления у числа знака "минус" (если он есть). Например, $|3| = 3$, $|-17| = 17$ и $|0| = 0$. Полезно запомнить такое простое правило: если значение t-статистики по абсолютной величине больше 2, то нулевую гипотезу отвергают, в противном случае принимают. Это правило применяют при n больше 40, используя число 2 как аппроксимацию t-значения 1,96. Таким образом, просмотрев колонку t-статистик, можно легко и быстро принять решение об их значимости. Например, числа 6,81; −4,97; 13,83; 2,46 и −5,81 — это значимые t-статистики, а числа 1,23; −0,51; 0,02; −1,86 и 0,75 — это незначимые t-статистики. (Отрицательное значение t-статистики говорит о том, что среднее значение выборки $\bar{X}$ меньше заданного значения μ_0.)

А что делать, если значение t-статистики *точно равно* t-значению из таблицы. Это имеет место, когда значение μ_0 точно совпадает с границей доверительного интервала. Как быть в таком случае? К счастью, это почти никогда не случается. Тем не менее вы можете увеличить точность, вычислив больше цифр после запятой, или же сделать вывод о том, что результат "значим, но является пограничным".

Несмотря на то что для решения вопроса о значимости значение t-статистики можно легко сравнить с числом 2 (или с более точным значением из t-таблицы), необходимо помнить, что значение t-статистики измеряется не в тех же единицах, что и исходные данные. Поскольку единицы измерения в числителе и знаменателе t-статистики взаимно сокращаются, результат является безразмерной величиной. Эта величина представляет собой расстояние между $\bar{X}$ и μ_0, выраженное в количестве *стандартных ошибок*, а не в долларах, милях на галлон, людях и других единицах, в которых измерены исходные данные.

Таблица 10.2.6. Использование t-статистики для проверки гипотезы

Если t-статистика меньше по абсолютной величине t-значения из t-таблицы ($\lvert t_{статистика}\rvert < \lvert t_{табл}\rvert$), то:
Принять нулевую гипотезу H_0 как допустимую возможность.
Не принимать альтернативную гипотезу H_1.
Выборочное среднее $\bar{X}$ *незначимо отличается* от заданного значения μ_0.
Наблюдаемая разница между выборочным средним $\bar{X}$ и заданным значением μ_0 может быть обусловлена только лишь случайностью.
Результат проверки *не* является *статистически значимым.*
(Все перечисленные выше утверждения эквивалентны.)
Если t-статистика больше по абсолютной величине t-значения из t-таблицы ($\lvert t_{статистика}\rvert > \lvert t_{табл}\rvert$), то:
Принять альтернативную гипотезу H_1.
Отклонить нулевую гипотезу H_0.
Выборочное среднее $\bar{X}$ *значимо отличается* от заданного значения μ_0.
Наблюдаемая разница между выборочным средним $\bar{X}$ и заданным значением μ_0 не может быть обусловлена только лишь случайностью.
Результат является *статистически значимым.*
(Все перечисленные выше утверждения эквивалентны.)

Кроме того, нет существенного различия в подходах к проверке гипотез с точки зрения доверительного интервала и t-статистики. Чтобы проверить это, рассмотрим еще раз предыдущие примеры.

В примере о добавке, "повышающей объем продукции", среднее выборки равно $\bar{X} = 39,6$ тонны, стандартная ошибка $S_{\bar{x}} = 4,2$ тонны, размер выборки $n = 7$ и заданное значение $\mu_0 = 32,1$ тонны. Заданное значение попадает в доверительный интервал, который находится между 29,3 и 49,9. Исходя из этого принимаем нулевую гипотезу. Если вместо этого вычислить t-статистику, получим:

$$t_{статистика} = \frac{\bar{X} - \mu_0}{S_{\bar{x}}} = \frac{39,6 - 32,1}{4,2} = 1,785714.$$

Поскольку абсолютное значение t-статистики 1,785714 меньше значения из t-таблицы 2,447, то нулевая гипотеза принимается. Таким образом, использование t-статистики дает тот же результат, что и использование доверительного интервала.

В примере о том, как влияет обладание акциями служащими компании на качество продукции, среднее значение оценки в выборке равно $\bar{X} = 0,35$, стандартная ошибка составляет $S_{\bar{x}} = 0,14$, размер выборки $n = 343$ и заданное значение $\mu_0 = 0$. Заданное значение *не* попадает в доверительный интервал, который находится между 0,08 и 0,62. Исходя из этого принимается альтернативная гипотеза. Если вместо построения доверительного интервала вычислить t-статистику, то получим следующее:

$$t_{статистика} = \frac{\bar{X} - \mu_0}{S_{\bar{x}}} = \frac{0,35 - 0}{0,14} = 2,50.$$

Ввиду того, что полученное значение t-статистики по абсолютной величине больше табличного значения, равного 1,960, принимается альтернативная гипотеза. И снова, как и должно быть, использование t-статистики дает тот же результат, что метод доверительного интервала.

В примере с биномиальным распределением, в котором рассматривается качество микросхем, в выборке размером $n = 500$ количество особо быстрых микросхем равно 58, биномиальная доля $p = 0,116$, стандартная ошибка $S_p = 0,0143$ и заданное значение $\pi_0 = 0,10$. Заданное значение принадлежит доверительному интервалу, который находится между 0,088 и 0,144. Исходя из этого принимают нулевую гипотезу. Если вместо построения доверительного интервала вычислить t-статистику, то получим следующее:

$$t_{статистика} = \frac{p - \pi_0}{S_p} = \frac{0,116 - 0,10}{0,0143} = 1,12.$$

Поскольку значение t-статистики, равное 1,12, по абсолютной величине меньше табличного значения, равного 1,960, принимают нулевую гипотезу, делая тот же вывод, что и при проверке гипотезы с использованием доверительного интервала.

10.3. Интерпретация проверки гипотезы

Теперь, когда вам известен механизм выполнения проверки гипотезы и обычные способы описания результата проверки, самое время узнать, какое вероятностное утверждение стоит за всем этим. Как и в случае доверительных интервалов, поскольку невозможно сделать вывод, корректный в 100% случаев, делают вывод о неизвестном среднем значении генеральной совокупности, корректный в 95% (или в 90%, или в 99%, или в 99,9%) случаев.

Обычно формальные детали процедуры проверки гипотезы определяют в терминах различных возможных *ошибок*, которые могут быть допущены. В результате проверки гипотезы на основании информации, полученной из выборочных данных, принимают одну из гипотез. Вы можете оказаться правы или не правы, поскольку гипотезы представляют собой утверждения о *генеральной совокупности*, о которой у вас нет полной информации. В общем, вы не можете быть полностью уверены в правильности своего выбора. Конечно, вы надеетесь, что вы правы, однако в зависимости от ситуации можете получить или не получить действительно убедительное вероятностное утверждение.

В основе каждого из типов ошибок лежат различные предположения относительно того, какая из гипотез *действительно* является верной. Конечно, в действительности вы обычно не можете знать, какая из гипотез верна, даже если вы уже решили принять одну из них. Однако, чтобы понять результаты проверки гипотезы, полезно посмотреть на них с точки зрения всех возможных исходов такой проверки.

Ошибки I и II рода

Если нулевая гипотеза в действительности является верной (хотя на самом деле вы никогда точно не знаете, так это или нет), но вы ошибочно решаете отвергнуть ее и принять альтернативную гипотезу, то вы совершаете ошибку I рода. Вероятность появления ошибки первого рода (когда истинна нулевая гипотеза) обычно ограничивается на уровне 5%:

$$P(\text{ошибка I рода при истинной нулевой гипотезе } H_0) = 0,05 .$$

Вероятность появления ошибки I рода можно контролировать, потому что нулевая гипотеза является очень определенной, и таким образом есть точное значение вероятности. Например, полагая, что нулевая гипотеза $H_0 : \mu = \mu_0$ верна, вы тем самым полагаете, что среднее значение генеральной совокупности известно. Ну а если вы знаете среднее генеральной совокупности, то вероятность можно легко вычислить.

Проверку на других уровнях (10, 1, 0,1%) можно выполнять, используя другие t-значения из t-таблицы, — например, работая с другими доверительными интервалами (90, 99 и 99,9% соответственно). Если вы не хотите ошибаться в 5% случаев, когда нулевая гипотеза верна, вы можете проводить проверку на уровне 1% (используя соответствующее t-значение из того столбца t-таблицы, в нижней части которого записано число 2,576) и тогда вероятность того, что вы совершите ошибку I рода (при условии, что нулевая гипотеза верна), будет равна всего лишь 1%.

Если в действительности верна альтернативная гипотеза (опять же, вы не знаете наверняка, так это или нет), но вы ошибочно решили принять нулевую гипотезу, то вы совершаете ошибку II рода. Вероятность ошибки II рода трудно контролировать:

$$P(\text{ошибка II рода при верной гипотезе } H_1) \text{ трудно контролировать.}$$

Контролировать вероятность ошибки II рода сложно, потому что эта вероятность изменяется в зависимости от истинного значения μ[8]. Допустим, что значение μ близко к значению μ_0. Тогда из-за случайности в данных эти два значения будет трудно разделить. Например, предположим, нулевая гипотеза утверждает, что μ равно 15,00000, а в действительности μ равно 15,00001. Тогда, несмотря на то, что нулевая гипотеза формально верна (так как $15,00000 \neq 15,00001$), практически эти значения разделить трудно и вероятность ошибки II рода будет приблизительно равна 95%. В то же время, если значение μ сильно отличается от 15, вероятность ошибки II рода будет почти равна нулю, что очень приятно. Таким образом, поскольку вероятность ошибки II рода сильно зависит от истинного значения μ, ее трудно контролировать. Оба типа ошибок показаны на рис. 10.3.1.

Истина	Ваше решение: Принять нулевую гипотезу	Ваше решение: Принять альтернативную гипотезу
Нулевая гипотеза	Правильное решение	Ошибка I рода (контролируемая на уровне 0,05 или любом другом уровне)
Альтернативная гипотеза	Ошибка II рода (трудноконтролируемая)	Правильное решение

Рис. 10.3.1. Решение о принятии гипотезы может быть верным, а может быть и неверным. В зависимости от того, какая из гипотез фактически является правильной, различают ошибки I и II рода. Легко контролировать только ошибку I рода, обычно это делают на уровне 5%

[8] В принципе, вероятность можно вычислить для каждого значения μ. Полученная таким образом таблица или график создают основу так называемой *мощности* теста. Это представляет собой анализ типа, *что если* дающий свойства ошибки II рода теста при каждом возможном значении μ.

Условия применимости

Для проверки статистических гипотез должны выполняться определенные условия. Поскольку проверку можно проводить на основе доверительных интервалов, условия применимости для проверки гипотез аналогичны условиям для построения доверительных интервалов. Для применимости проверки статистических гипотез необходимо выполнение следующих условий: (1) набор данных является случайной выборкой из рассматриваемой генеральной совокупности, (2) либо измеряемые величины приблизительно нормально распределены, либо размер выборки настолько велик, что в соответствии с центральной предельной теоремой выборочное среднее распределено приблизительно нормально.

Что произойдет, если эти условия не будут выполнены? Рассмотрим вероятность ошибки первого рода (ошибочное отклонение нулевой гипотезы, которая в действительности является верной). Вероятность этой ошибки уже не будет контролироваться на низком и управляемом уровне 5% (или другом заявленном вами уровне). Вместо этого истинная вероятность ошибки может быть намного выше или ниже 5%. В этом случае определение значимости уже не так важно, поскольку событие "неверно определенная значимость" встречается все более часто.

Если набор данных не является случайной выборкой из рассматриваемой генеральной совокупности, то статистика ничего не может сделать для вас, поскольку данные просто не содержат необходимой информации.

Предположим, что данные представляют собой случайную выборку, но распределение не является нормальным. Если распределение ваших данных действительно сильно отличается от нормального, можно попытаться преобразовать данные (например, если все значения положительны, применить логарифмирование), чтобы получить распределение, близкое к нормальному. Обратите внимание, что после такого преобразования данных вы будете проверять гипотезу не для среднего значений генеральной совокупности, а для среднего *логарифмов* значений генеральной совокупности. Другое решение состоит в использовании непараметрических тестов, которые будут описаны в главе 16.

Гипотезы не могут быть вероятно истинными или вероятно ложными

Возможно, вы заметили, что мы никогда не говорим, что гипотеза "вероятно" истинна или ложна. Мы всегда очень осторожно принимаем или отклоняем гипотезу, вынося при этом определенное, точное решение. Мы говорим об ошибках, которые мы можем совершить, и о вероятностях *этих ошибок*, но не о вероятности того, истинной или ложной является гипотеза. Причина проста. *В гипотезе нет ничего случайного!*

Нулевая гипотеза или истинна, или ложна в зависимости от значения среднего генеральной совокупности μ. Среднее генеральной совокупности не содержит никакой случайности. Аналогично альтернативная гипотеза или истинна, или ложна, и несмотря на то, что вы не знаете точно, истинна она или ложна, гипотеза не содержит случайности. Случайность является следствием случайного процесса осуществления выборки, который дает нам данные для принятия решения.

Таким образом, ваше *решение* (о выборе гипотезы) известно и случайно, точно так же, как и выборочная статистика, поскольку это решение основано на данных. В то же время истинная гипотеза определена, но неизвестна точно так же, как параметр генеральной совокупности.

Статистическая значимость и уровни проверки

Принято считать, что результат является статистически значимым, если альтернативная гипотеза принимается на уровне 5% (например, с использованием стандартного 95% доверительного интервала). Обратите внимание, что здесь слово "значимый" употребляется не в общепринятом смысле. Обычно слово "значимый" означает "особенно важный". А в статистике это не так.

Хорошей иллюстрацией этого служит случай из жизни. Однажды ко мне пришел адвокат, обеспокоенный тем, что противная сторона, подавшая иск, обнаружила *статистически значимую разницу* между размером двери, которая имела отношение к несчастному случаю, и другими подобными дверями в этом же здании. Но после того как адвокату пояснили специальный статистический смысл слова *значимый*, он успокоился, поскольку понял, что противная сторона *не* показала, что данная дверь значительно отличается от других дверей. Они лишь продемонстрировали, что имело место *статистически фиксируемое* различие. В действительности различие было небольшим. Но при тщательном измерении его можно было зафиксировать! Это было не случайное отличие от других дверей (нулевая гипотеза), это отличие было систематическим (альтернативная гипотеза). Однако, хотя разница была статистически значимой, она не была настолько велика, чтобы иметь большое значение. Это как со снежинками; все они действительно отличаются друг от друга, но, по существу, все одинаковы.

Мораль этой истории в том, что на вас не должно автоматически производить большое впечатление, если кто-то похвастается "значимым результатом". Если слово "значимый" используют в статистическом смысле, то это говорит только о том, что исключена случайность. Необходимо еще проанализировать данные, чтобы определить, достаточно ли сильно влияние, чтобы иметь для вас значение. Статистические методы оперируют только числами. Чтобы принять решение о важности и значимости статистических результатов, вы должны использовать свои знания из других областей.

Существует еще одна причина, по которой не стоит сильно увлекаться статистически значимыми результатами. В течение всей вашей жизни приблизительно 5% всех результатов проверки гипотез *для ситуаций, в которых нулевая гипотеза действительно истинна*, будут значимыми. Это подразумевает, что приблизительно одна из каждых 20 неинтересных ситуаций будет *ошибочно* объявлена значимой (т.е. будет допущена ошибка I рода). Один специалист в области лекарственных средств однажды заметил, что около 5% проверяемых лекарств для лечения одной тяжелой болезни имели значимый лечебный эффект. Поскольку эта доля близка к доле тех лекарств, которые могли быть отнесены к эффективным *по ошибке, даже если ни одно из них не было фактически эффективным*, то это замечание свидетельствует о неприемлемости всей этой программы поиска эффективных лекарств.

Используя значения из различных столбцов t-таблицы, можно проверить гипотезу на уровнях 10, 5, 1 и 0,1%. Уровень проверки, или уровень значимости, представляет собой вероятность совершить ошибку I рода при условии, что нулевая гипотеза является истинной.[9] Когда вы отвергаете нулевую гипотезу и принимаете альтернативную, то вы можете утверждать, что ваш результат является *значимым* на уровне 10, 5, 1 или 0,1%, — в зависимости от того, какой столбец t-таблицы вы использовали. Чем ниже уровень, на котором вы можете определить значимость, тем более впечатляющим будет результат. Например, получение результата, значимого на уровне 1%, является более впечатляющим, чем получение значимости на уровне 10 или 5%; вероятность ошибки первого рода в таком случае меньше, и аргументы против нулевой гипотезы сильнее. Обычно используют следующие фразы для описания результатов.

Незначимый	Отсутствие значимости на обычном уровне 5%
Значимый	Значимость на обычном уровне 5%
Высоко значимый	Значимость на уровне 1%
Очень высоко значимый	Значимость на уровне 0,1%

Что делать, если вам удалось определить значимость на нескольких уровнях? Следует радоваться! Серьезно, чем ниже уровень, на котором вам удалось определить значимость, тем сильнее доказательство против нулевой гипотезы. Поэтому следует указывать *наименьший* из уровней значимости. Например, если найдена значимость на уровнях 5 и 1%, достаточно указать, что ваш результат является высоко значимым (т.е. значим на уровне 1%).

Если значимость найдена на некотором уровне, то значимость обязательно имеет место на всех *более высоких* уровнях.[10] Таким образом, высоко значимый результат (значимый на уровне 1%) должен *обязательно* (и это можно строго доказать математически) быть значимым на уровнях 5 и 10%. Однако он может быть, а может и не быть значимым на уровне 0,1%.

Доверительная вероятность (p-значение)

Результат проверки статистической гипотезы характеризуется также значением доверительной вероятности (p-значением), которое показывает вероятность того, что данные соответствуют нулевой гипотезе. Малые значения свидетельствуют об удивительности такого события и приводят к тому, что H_0 отклоняют. Обычно H_0 отвергают, когда p-значения меньше 0,05. В предположении о том, что нулевая гипотеза является верной, p-значение равно вероятности наблюдать имеющиеся данные (или даже данные, еще в большей мере не соответствующие H_0). Если H_0 является верной, то появление малых p-значений маловероятно, поэтому малые p-значения приводят к тому, что H_0 отклоняют. Например, если

[9] В более общих случаях, включая односторонний направленный тест, уровень проверки, или уровень значимости, определяется более тщательно как *максимальная* вероятность совершить ошибку I рода, максимизированная с учетом всех возможностей, включенных в нулевую гипотезу.

[10] Обратите внимание, что *больший* уровень значимости фактически означает *менее* впечатляющий результат. Например, результат при 1% уровне значимости является высоко значимым, а при уровне значимости 5% (большем) — только значимым.

$p = 0,001$, то данные с большими отличиями от H_0 встречаются реже, чем один раз на тысячу случайных выборок. Вместо того чтобы предполагать, что это редкое (одно на тысячу попыток) событие может произойти (потому что оно происходит по крайней мере не чаще), проще объяснить, что нулевая гипотеза H_0 является ложной и должна быть отклонена. Обычно p-значения описывают следующим образом.

Описание	Интерпретация
Незначимый ($p > 0,05$)	Незначимый на обычном уровне 5%
Значимый ($p < 0,05$)	Значимый на обычном уровне 5%, но незначимый на уровне 1%
Высоко значимый ($p < 0,01$)	Является значимым на уровне 1%, но незначимым на уровне 0,1%
Очень высоко значимый ($p < 0,001$)	Значимый на уровне 0,1%

В некоторых областях исследования можно рассматривать и результаты, значимые на уровне 10%. Это означает, что вероятность совершить ошибку, отклоняя фактически правильную нулевую гипотезу, равна 0,1. Во многих областях исследования такой высокий уровень ошибки недопустим. Однако в некоторых сферах непредсказуемость и изменчивость данных создает трудности для получения результата на (обычном) уровне значимости 5%. Если вы работаете именно в такой области, используйте в качестве возможной следующую формулировку утверждения о p-значении.

> Результат является значимым на уровне 10%, но не на обычном уровне 5% ($p < 0,1$).

Часто p-значения вставляют непосредственно в текст, например: "Обнаружено, что стиль музыки оказывает значимое влияние ($p < 0,05$) на поведение покупателей". Также p-значения включают либо в текст, либо в таблицу в виде сноски, как, например: "В результате выполнения новой программы производительность значимо[11] увеличилась".

Большинство современных статистических пакетов программ в качестве результата проверки статистической гипотезы вычисляет точное p-значение. Если проверка осуществляется на уровне 5% и p-значение представляет собой любое число меньше 0,05, то результат будет значимым (например, $p = 0,0358$ соответствует значимому результату, а $p = 0,2083$ — незначимому). Учтите, что p-значение представляет собой статистику (а не параметр генеральной совокупности), потому что его вычисляют на основе данных выборки (и заданного опорного значения).

Рассмотрим пример выполнения проверки относительно того, значимо ли среднее значение объема продукции (переменная YIELD) $\bar{X} = 39,6$ отличается от заданного значения $\mu_0 = 32,1$ тонны (исходя из $n = 7$ наблюдений и стандартной ошибки $S_{\bar{X}} = 4,2$). Компьютер может выводить результаты вычисления в таком виде.

[11] ($p < 0,05$).

	N	Среднее (MEAN)	Стандартное отклонение (STDEV)	Стандартная ошибка (SE MEAN)	T (T)	p-значение (p-VALUE)
Объем (YIELD)	7	39,629	11,120	4,203	1,79	0,12

Поскольку вычисленное значение ($p = 0{,}12$) больше обычного уровня проверки 5% (т.е. 0,12 > 0,05), то результат "незначимый ($p > 0{,}05$)". Этот результат (об отсутствии значимости) можно также получить, сравнив вычисленное значение t-статистики (1,79) с табличным t-значением (из t-таблицы). Точное p-значение в данном случае можно интерпретировать таким образом: в предположении, что среднее генеральной совокупности равно заданному значению $\mu_0 = 32{,}1$, вероятность таких больших различий (между наблюдаемым средним и заданным значением) равна 12%. Как правило, вероятность 12% не рассматривают как выходящую за рамки обычного, но события с вероятностью не более 5% считают маловероятными. С другой стороны, можно потребовать у компьютера вычислить 95% доверительный интервал.

	N	Среднее (MEAN)	Стандартное отклонение (STDEV)	Стандартная ошибка (SE MEAN)	95% доверительный интервал (95,0 PERCENT C.I.)
Объем (YIELD)	7	39,63	11,12	4,20	(29,34;49,92)

Здесь видно, что результат проверки не является значимым, так как заданное значение $\mu_0 = 32{,}1$ находится за пределами доверительного интервала (от 29,34 до 49,92).

Далее рассмотрим пример проверки того, считают ли менеджеры, что существует взаимосвязь между тем, что служащие владеют акциями компании, и повышением качества продукции (переменная SCORE). Компьютер может представить результаты своих вычислений следующим образом.

	N	Среднее (MEAN)	Стандартное отклонение (STDEV)	Стандартная ошибка (SE MEAN)	T (T)	p-значение (p-VALUE)
Качество (SCORE)	343	0,350	2,593	0,140	2,50	0,013

Здесь мы видим, что p-значение равно $p = 0{,}013$, поэтому результат является значимым на уровне 5% (поскольку $p < 0{,}05$), но не является значимым на уровне 1% (поскольку $p > 0{,}01$).

Заметим, что в случае биномиального распределения существуют две различные величины, обозначаемые обычно одной и той же буквой p. Одна из них представляет собой наблюдаемую частоту в выборке $p = X/n$, а другая — p-значение, вычисленное при проверке статистической гипотезы относительно конкретно заданного опорного значения.

10.4. Односторонняя проверка

Все проверки, которые мы до сих пор рассматривали, являются двусторонними, так как они проверяют нулевую гипотезу $H_0 : \mu = \mu_0$ против альтернативной гипотезы $H_0 : \mu \neq \mu_0$. Эта альтернативная гипотеза является двусторонней, потому что среднее совокупности может быть как больше, так и меньше заданного опорного значения μ_0.

Однако вас может не интересовать проверка того, *отличается* ли среднее генеральной совокупности от заданного значения. Вас может интересовать более конкретный вопрос: является ли среднее генеральной совокупности *больше* (в одних случаях) или *меньше* (в других случаях) заданного значения. Например, вы купите систему только в том случае, если возможная экономия в долгосрочном плане окажется *значимо выше* некоторого конкретного числа (заданное значение μ_0). Или вас может интересовать возможность заявить о высоком качестве вашей продукции, поскольку объем брака *значимо меньше* некоторого существенно малого числа.

Необязательно использовать одностороннюю проверку, чтобы утверждать, что среднее значение выборки значимо больше или значимо меньше заданного значения. Для этой цели можно использовать и двустороннюю проверку. Если результат двусторонней проверки оказывается значимым (т.е. альтернативная гипотеза принимается), то с учетом того, больше или меньше среднее выборки $\bar{X}$, чем заданное опорное значение, можно сделать следующие выводы о значимости.

Использование двусторонней проверки для односторонних заключений[12]

Если двусторонняя проверка значима и $\bar{X} > \mu_0$	Среднее значение выборки $\bar{X}$ значимо больше заданного значения μ_0
Если двусторонняя проверка значима и $\bar{X} < \mu_0$	Среднее значение выборки $\bar{X}$ значимо меньше заданного значения μ_0

Однако бывает выгодно использовать именно одностороннюю проверку. Соблюдая все требования, можно получить, что результат односторонней проверки значим, в то время как результат двусторонней проверки незначим. Почему это возможно? Концентрируя внимание только на одной стороне (на одном направлении) и игнорируя другую, односторонняя проверка может лучше определить различие между средним и заданным опорным значением именно на этой стороне. Платой за это будет то, что односторонняя проверка неспособна определить различие (независимо от того, насколько оно велико) на другой стороне (в другом направлении).

В процедуре одностороннего t-теста нулевая гипотеза утверждает, что значение μ находится по одну сторону от μ_0, а альтернативная гипотеза утверждает, что значение μ находится по другую сторону от μ_0. (Случай $\mu = \mu_0$ всегда включают в нулевую гипотезу, которую необходимо опровергнуть. Это гарантирует, что принятие альтернативной гипотезы на некотором уровне значимости позволяет сделать более строгое заключение: либо "значимо больше, чем" либо "зна-

[12] Помните, что *двустороннее* заключение звучало бы как "$\bar{X}$ значимо отличается от μ_0".

чимо меньше, чем"[13]). Гипотезы для двух видов односторонней (направленной) проверки формулируются следующим образом.

Односторонняя проверка того, что μ меньше μ_0

$$H_0: \mu \geq \mu_0$$

Нулевая гипотеза утверждает, что неизвестное среднее значение генеральной совокупности *по меньшей мере так же велико*, как известное заданное значение μ_0.

$$H_1: \mu < \mu_0$$

Альтернативная гипотеза утверждает, что неизвестное среднее значение генеральной совокупности *меньше*, чем известное заданное значение μ_0.

Односторонняя проверка того, что μ больше μ_0

$$H_0: \mu \leq \mu_0$$

Нулевая гипотеза утверждает, что неизвестное среднее значение генеральной совокупности *не больше*, чем известное заданное значение μ_0.

$$H_1: \mu > \mu_0$$

Альтернативная гипотеза утверждает, что неизвестное среднее значение генеральной совокупности *больше*, чем известное заданное значение μ_0.

Существует важное условие, выполнение которого обязательно для использования односторонней проверки статистической гипотезы. По сути, это то же условие, выполнение которого требуется и при построении одностороннего доверительного интервала.

Чтобы использовать односторонний тест, следует быть уверенным, что *независимо от того, как ведут себя данные*, односторонний тест будет продолжать применяться на той же стороне ("больше, чем" или "меньше, чем"). Если в результате изменения характера данных возникает желание использовать тест *на другой стороне* вместо ранее планируемой, то необходимо перейти к использованию двустороннего теста. Если существуют сомнения, следует также использовать двусторонний тест.

В частности, использование одностороннего теста оставляет место для критики. Поскольку интересующее вас решение может быть субъективным, то желание сосредоточиться на том, что интересует непосредственно вас, может не соответствовать мнению тех, кого вы хотите убедить. Если необходимо убедить людей, которые могут иметь другую точку зрения (например, проверяющих или противостоящих вам юристов), следует использовать двусторонний тест и делать на его основе одностороннее (направленное) заключение. С другой стороны, если необходимо убедить "дружески" настроенных людей с интересами, аналогичными вашим (например, сотрудников вашего отдела или вашей фирмы), то в случае выполнения указанного ранее условия можно использовать преимущества одностороннего теста.

[13] С технической точки зрения случай равенства может быть включен в исследуемую альтернативную гипотезу. При этом процедура проверки будет такой же и значимость будет определяться точно так же, но вывод будет более слабым.

Альтернативная гипотеза будет принята только тогда, когда есть убедительные доказательства против нулевой гипотезы. Альтернативную гипотезу принимают тогда, когда среднее выборки $\bar{X}$ и заданное значение μ_0 связаны таким соотношением, которое описано в альтернативной гипотезе, *и* при этом указанные значения достаточно сильно различаются (находятся на расстоянии $t_{табл.}$ или больше значений стандартной ошибки, которая представляет собой допустимое отклонение от среднего значения). Существуют два способа выполнения односторонней проверки: с использованием доверительного интервала и с использованием t-статистики.

Пример. *Запуская новый продукт...*

Допустим, что анализ себестоимости нового потребительского продукта предполагает, что продукт будет успешным тогда, когда больше 23% потребителей захотят его попробовать. Эти 23% представляют собой заданное значение μ_0, которое получено путем теоретического анализа, не на основе данных случайной выборки. Чтобы решить вопрос о запуске продукта в производство, вы собираете информацию у случайно отобранных потребителей и вычисляете односторонний доверительный интервал. Исходя из собранных данных вы ожидаете, что 44,1% потребителей захотят испытать новый продукт ($\bar{X} = 44{,}1\%$), и ваше утверждение относительно доверительного интервала гласит о том, что вы на 95% уверены, что *по меньшей мере* 38,4% потребителей захотят испытать новый продукт. Поскольку заданное значение $\mu_0 = 23\%$, характеризующее безубыточность, находится вне доверительного интервала (и, следовательно, *неприемлемо* предполагать, что среднее значение равно 23%), у вас есть убедительное доказательство, что среднее в генеральной совокупности больше 23%. Эта ситуация кратко представлена в таблице 10.4.1.[14]

Решение состоит в том, чтобы принять альтернативную гипотезу H_1, поскольку заданное значение находится за пределами доверительного интервала (т.е. 23% не означает "по крайней мере 38,4%").

Поскольку заданное значение равно 23%, что намного меньше нижней границы доверительного интервала, можно попытаться рассмотреть более впечатляющий уровень значимости. Действительно, используя значение 3,090 из t-таблицы для построения одностороннего доверительного интервала, можно утверждать с доверительной вероятностью 99,9%, что среднее совокупности будет не меньше 33,4%. Так как заданное значение 23% находится даже вне этого доверительного интервала, результат можно охарактеризовать как *очень высоко значимый* ($p < 0{,}001$).

Как выполнять проверку

В табл. 10.4.2 показано, как выполнить односторонний тест. Таблица содержит подробные инструкции для обоих типов проверок (т.е. проверок того, что $\bar{X}$ значимо больше или значимо меньше μ_0) с использованием как метода доверительного интервала, так и метода t-статистики. Приведены также оба типа возможных выводов (значимо или нет) и их интерпретация.

Полезно руководствоваться таким принципом: результат проверки значим тогда, когда заданное значение μ_0 не попадает в односторонний доверительный интервал, построенный в направлении, соответствующем альтернативной гипотезе.

При использовании метода доверительного интервала необходимо помнить, что есть два различных односторонних доверительных интервала. Вам нужно выбрать тот, который относится к стороне, относительно которой сформулирова-

[14] Поскольку имеет место биномиальное распределение, $\bar{X}$ можно заменить на p, $S_{\bar{X}}$ — на S_p, μ — на π и μ_0 — на π_0. Чтобы показать, что этот пример корректен в том виде, в каком он представлен здесь, отметим, что $\bar{X} = p$ для набора данных $X_1, \ldots, X_n$, где каждый из элементов данных равен либо 0, либо 1, в зависимости от ответа каждого из потребителей.

Таблица 10.4.1. Тест для процента потребителей, желающих испытать новый продукт (с использованием доверительного интервала)

Нулевая гипотеза	$H_0 : \mu \le \mu_0$	$H_0 : \mu \le 23\%$
Альтернативная гипотеза	$H_1 : \mu > \mu_0$	$H_1 : \mu > 23\%$
Среднее	$\bar{X}$	44,1%
Стандартная ошибка	$S_{\bar{x}}$	3,47%
Размер выборки	n	205
Заданное значение	μ_0	23%
Доверительный интервал	$\bar{X} - t_{табл} S_{\bar{x}}$	"Мы на 95% уверены, что среднее генеральной совокупности составляет по меньшей мере 38,4%"
Решение	Принять альтернативную гипотезу H_1	"Мы ожидаем, что значимо больше, чем 23% покупателей попробуют наш продукт"*

*Значимо ($p < 0,05$) для односторонней проверки.

на альтернативная гипотеза. Например, если исследуемая альтернативная гипотеза имеет вид неравенства $H_1 : \mu > \mu_0$, то необходимый вам односторонний доверительный интервал будет состоять из всех значений μ, которые *по крайней мере не меньше* соответствующего числа, вычисленного по формуле $\bar{X} - t_{табл} S_{\bar{x}}$ с использованием t-значения для одностороннего интервала, взятого из t-таблицы.

На рис. 10.4.1 показано, что для принятия решения о том, что $\bar{X}$ значимо больше μ_0, расстояние между ними должно быть достаточно большим, чтобы эту разницу нельзя было бы объяснить только случайностью.

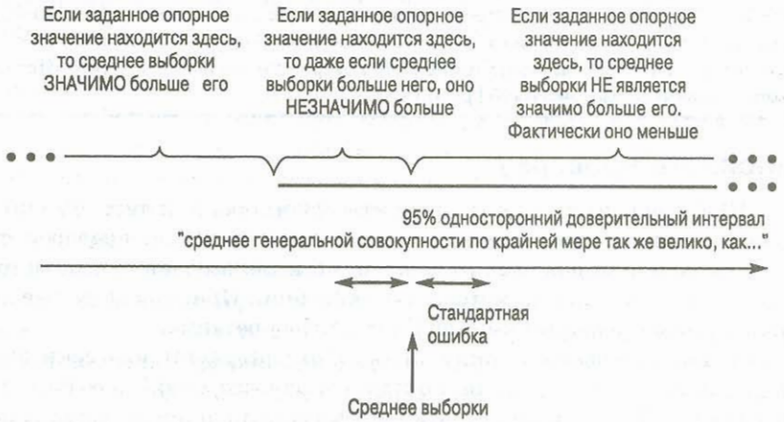


Рис. 10.4.1. Использование односторонней проверки для принятия решения о том, действительно ли μ больше, чем заданное значение μ_0. Односторонний доверительный интервал использует ту же сторону (то же направление), что и альтернативная гипотеза (т.е. именно те разумно допустимые значения μ, которые не меньше граничного значения интервала). Только если заданное значение μ_0 находится намного ниже среднего выборки, можно принять решение о том, что среднее выборки значимо больше

Таблица 10.4.2. Односторонняя проверка того, что μ больше μ_0

Проверяется нулевая гипотеза $H_0 : \mu \le \mu_0$ против альтернативной гипотезы $H_1 : \mu > \mu_0$.

Утверждение о доверительном интервале: "Мы на 95% уверены, что среднее генеральной совокупности по крайней мере не меньше, чем $\bar{X} - t_{табл} S_{\bar{x}}$".

t-статистика определяется следующим образом: $t_{статистика} = (\bar{X} - \mu_0) / S_{\bar{x}}$ (обратите внимание, что при односторонней проверке не используют абсолютное значение).

Действительно ли $\bar{X} - t_{табл} S_{\bar{x}} \le \mu_0$? В рамках метода доверительного интервала формулируем вопрос: находится ли заданное значение μ_0 внутри доверительного интервала? В рамках метода t-статистики формулируем эквивалентный вопрос: выполняется ли неравенство $t_{статистика} \le t_{табл}$? Если да, то:

Принять нулевую гипотезу H_0 как приемлемую возможность.

Не принимать альтернативную гипотезу H_1.

Среднее выборки $\bar{X}$ *не является значимо большим*, чем заданное значение μ_0.

Если $\bar{X}$ больше μ_0, то наблюдаемое различие можно приемлемо объяснить как возникшее только благодаря случайности.

Результат *не является статистически значимым.*

Действительно ли $\bar{X} - t_{табл} S_{\bar{x}} > \mu_0$? В рамках метод доверительного интервала формулируем вопрос: находится ли заданное значение μ_0 вне доверительного интервала? В рамках метода t-статистики формулируем эквивалентный вопрос: выполняется ли неравенство $t_{статистика} > t_{табл}$? Если, да то:

Принять альтернативную гипотезу H_1.

Отклонить нулевую гипотезу H_0.

Среднее выборки $\bar{X}$ *значимо больше*, чем заданное значение μ_0.

Наблюдаемое различие между средним выборки $\bar{X}$ и заданным значением μ_0 *нельзя* объяснить только случайностью.

Результат является *статистически значимым.*

Односторонняя проверка того, что μ меньше μ_0

Проверяется нулевая гипотеза $H_0 : \mu \ge \mu_0$ против альтернативной гипотезы $H_1 : \mu < \mu_0$.

Утверждение о доверительном интервале: "Мы на 95% уверены, что среднее совокупности *не больше*, чем $\bar{X} + t_{табл} S_{\bar{x}}$".

t-статистика равна $t_{статистика} = (\bar{X} - \mu_0) / S_{\bar{x}}$ (обратите внимание, что при односторонней проверке не используют абсолютное значение).

Действительно ли $\bar{X} + t_{табл} S_{\bar{x}} \ge \mu_0$? В рамках метода доверительного интервала формулируем вопрос: находится ли заданное значение μ_0 внутри доверительного интервала? В рамках метода t-статистики формулируем эквивалентный вопрос: выполняется ли неравенство $t_{статистика} \ge -t_{табл}$? Если да, то:

Принять нулевую гипотезу H_0 как приемлемую возможность.

Отклонить альтернативную гипотезу H_1.

Среднее выборки $\bar{X}$ *незначимо меньше*, чем заданное значение μ_0.

Если $\bar{X}$ меньше, чем μ_0, то наблюдаемое различие можно объяснить как возникшее только благодаря случайности.

Результат *не является статистически значимым.*

Действительно ли $\bar{X} + t_{табл} S_{\bar{x}} < \mu_0$? В рамках метода доверительного интервала формулируем вопрос: находится ли заданное значение μ_0 вне доверительного интервала? В рамках метода t-статистики формулируем эквивалентный вопрос: выполняется ли неравенство $t_{статистика} < -t_{табл}$? Если, да то:

Принять исследуемую альтернативную гипотезу H_1.

Отклонить нулевую гипотезу H_0.

Среднее выборки $\bar{X}$ *значимо меньше*, чем заданное значение μ_0.

Наблюдаемое различие между средним выборки $\bar{X}$ и заданным значением μ_0 *нельзя* объяснить лишь случайностью.

Результат является *статистически значимым.*

На рис. 10.4.2 приведена соответствующая картинка для односторонней проверки на другой стороне (в другом направлении).

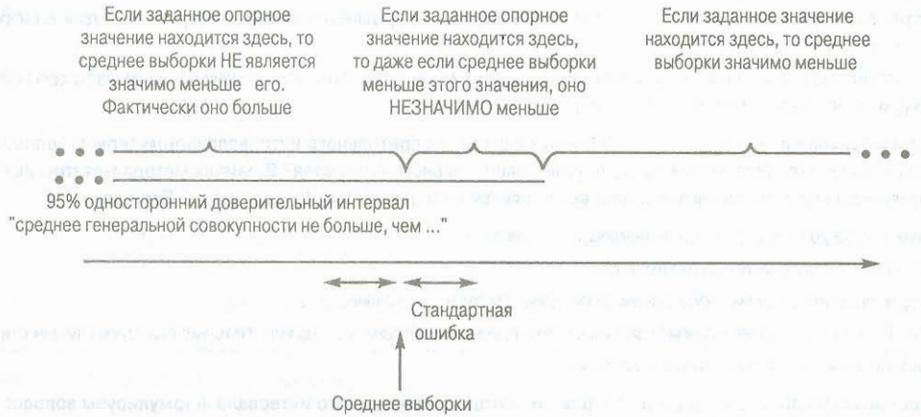


Рис. 10.4.2. Использование односторонней проверки для принятия решения о том, действительно ли μ меньше, чем заданное значение μ_0. Односторонний доверительный интервал использует ту же сторону (то же направление), что и альтернативная гипотеза (т.е. именно те приемлемо возможные значения μ, которые меньше или равны граничному значению интервала). Только если заданное значение μ_0 находится намного выше среднего выборки, можно принять решение о том, что среднее выборки значимо меньше

При использовании t-статистики проверку выполняют, сравнивая $t_{статистика} = \left(\bar{X} - \mu_0\right) / S_{\bar{X}}$ либо с табличным значением $t_{табл}$, либо с отрицательным значением $-t_{табл}$, в зависимости от того, с какой стороны (в каком направлении) производится проверка (точнее, в зависимости от формулировки альтернативной гипотезы: $H_1 : \mu > \mu_0$ или $H_1 : \mu < \mu_0$). Суть в том, что результат проверки является значимым, если данные соответствуют стороне альтернативной гипотезы и значение t-статистики велико по абсолютной величине (т.е. различие между $\bar{X}$ и μ_0 настолько велико, что его нельзя объяснить только случайностью). Обратите внимание, что t-статистика одна и та же для одно- и двусторонней проверки, но для принятия решения относительно значимости ее используют по-разному.

Пример. *Запуск нового продукта (пересмотренный вариант)*

Выше был рассмотрен пример запуска в производство нового потребительского продукта. Утверждалось, что запуск принесет прибыль только тогда, когда более 23% потребителей попробуют этот продукт. Соответствующие данные и результаты приведены в табл. 10.4.3. Теперь выполним такую же, как и ранее, одностороннюю проверку, но с использованием метода t-статистики.

Выносим решение принять альтернативную гипотезу H_1, поскольку $t_{статистика} > t_{табл}$, т.е. мы имеем 6,08 > 1,645 с использованием соответствующего критерия из табл. 10.4.2 для нашей альтернативной гипотезы ($H_1 : \mu > \mu_0$). К тому же, результат проверки гипотезы остается тем же (т.е. значимым), независимо от того, использован метод доверительного интервала или метод t-статистики. Фактически, поскольку значение t-статистики превышает табличное значение для одностороннего интервала (3,090) и для уровня 0,001, можно утверждать, что результат является *очень высоко значимым* ($p < 0,001$).

Таблица 10.4.3. Тест для процента потребителей, желающих испытать новый продукт (с использованием метода t-статистики)

Нулевая гипотеза	$H_0 : \mu \le \mu_0$	$H_0 : \mu \le 23\%$
Альтернативная гипотеза	$H_1 : \mu > \mu_0$	$H_1 : \mu > 23\%$
Среднее	$\bar{X}$	44,1%
Стандартная ошибка	$S_{\bar{X}}$	3,47%
Опорная величина	μ_0	23%
t-статистика	$t_{статистика} = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}$	$\frac{44,1 - 23}{3,47} = 6,08$
Критическое значение	$t_{табл}$	1,645
Решение	Принять альтернативную гипотезу H_1	"Мы ожидаем, что значимо больше, чем 23% потребителей, попробуют наш продукт"*

*Значимо ($p < 0,05$) для односторонней проверки.

Пример. *Сокращаются ли расходы*

Вы проверяете новую систему, которая предположительно снижает *переменные издержки производства*, или *стоимость производства изделия* (т.е. стоимость производства каждого дополнительного изделия за вычетом постоянных издержек, таких как арендная плата, которая не зависит от количества выпускаемых изделий). Ввиду того что внедрение новой системы будет сопровождаться дополнительными расходами, вам хотелось бы использовать ее только при наличии уверенности в том, что переменные издержки составляют менее $6,27 на единицу выпускаемой продукции.

Внимательно изучив 30 случайно выбранных изделий, изготовленных с использованием новой системы, вы обнаружили, что среднее значение переменных издержек производства составляет $6,05. Похоже, что переменные издержки в среднем меньше, чем целевое значение, равное $6,27. Но *значимо* ли оно меньше? Иными словами, можно ли ожидать, что в долгосрочном периоде средняя стоимость единицы продукции будет менее $6,27 или это случайная удача для этих 30 проверенных вами изделий? Исходя из имеющейся информации утверждать это нельзя, так как неизвестно, насколько случаен этот процесс. Действительно ли $6,05 меньше $6,27? Да, конечно. *Значимо* ли $6,05 меньше $6,27? На этот вопрос можно ответить, только сравнив эту разницу с разницей, обусловленной случайностью процесса, с использованием стандартной ошибки и табличного t-значения.

Итак, находим стандартное отклонение и вычисляем стандартную ошибку, которая равна $0,12. В табл. 10.4.4 подытожен результат односторонней проверки (показаны оба метода) того, значимо ли меньше требуемой величины ваши издержки.

Использование доверительного интервала дает вам 95% уверенность, что среднее значение переменных издержек меньше $6,25. Это значит, что вы можете еще в большей степени быть уверены, что это значение издержек меньше, чем заданное требуемое значение $6,27. Следовательно, результат является значимым. Можно также просто сказать, что заданное значение ($6,27) находится вне доверительного интервала, который имеет границу $6,25.

Используя метод t-статистики мы также получаем, что результат проверки значим, поскольку $t_{статистика} < -t_{табл}$, т.е. –1,833 < –1,699 с использованием соответствующего критического значения из табл. 10.4.2 для альтернативной гипотезы ($H_1 : \mu < \mu_0$).

В этом случае лучше использовать одностороннюю (направленную) проверку, потому что вас в действительности интересует только одно направление. Если вы сможете получить убедительное доказательство, что среднее значение переменных издержек меньше $6,27, то систему стоит применить. Если нет, то она вас не интересует. Используя одностороннюю проверку, вы соглашаетесь с тем, что если система дейст-

Таблица 10.4.4. Проверка для величины переменных издержек за длительный период

Заданное значение	μ_0	\$6,27
Нулевая гипотеза	$H_0: \mu \geq \mu_0$	$H_0: \mu \geq \$6,27$
Альтернативная гипотеза	$H_1: \mu < \mu_0$	$H_1: \mu < \$6,27$
Среднее	$\bar{X}$	\$6,05
Стандартная ошибка	$S_{\bar{X}}$	\$0,12
Размер выборки	n	30
Доверительный интервал	$\bar{X} + t_{табл} S_{\bar{X}}$	"Мы на 95% уверены, что среднее значение переменных издержек для долгосрочного периода меньше \$6,25"
t-статистика	$t_{статистика} = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}$	$\frac{6,05 - 6,27}{0,12} = -1,833$
Критическое значение	$-t_{табл}$	−1,699
Решение	Принять альтернативную гипотезу H_1	"Размер переменных издержек при использовании новой системы значимо меньше \$6,27"*

*Значимо ($p < 0,05$) для односторонней проверки.

вительно плоха, вы не сможете сказать, что "переменные издержки значимо *больше, чем...*" , вы сможете только сказать, что "переменные издержки *не являются значимо меньше*".

Если бы вы использовали двустороннюю проверку, которую также можно применить в этом случае, но которая является менее эффективной, вы обнаружили бы, что результат *не является значимым*! Двусторонний доверительный интервал простирается от \$5,80 до \$6,30 и *включает* заданное значение. Значение t-статистики остается тем же — 1,833, но t-значение для двустороннего интервала равно 2,045, а значит, больше абсолютной величины значения t-статистики (1,833). Таким образом, этот пример показывает, что результат односторонней проверки может быть значимым и в то же время результат соответствующей двусторонней проверки может быть незначимым. Это наблюдается только тогда, когда заданное значение (с которым производится сравнение) находится почти на границе одностороннего доверительного интервала, как в нашем примере.

Следует ли покупать систему? Это стратегический вопрос бизнеса, а не статистики. Используйте результаты проверки гипотезы в качестве исходных данных, но примите во внимание и другие факторы, такие как наличие инвестиционного капитала, необходимого персонала и взаимосвязь с другими проектами. Также не следует забывать, что хотя в результате проверки гипотезы вы приняли альтернативную гипотезу о том, что переменные издержки меньше порогового значения, этот результат *не* является абсолютно доказанным — остается место для ошибки. Вы не можете также сказать, какова вероятность того, что ваше решение неверно, потому что вы не знаете, какая из гипотез в действительности является верной. Наибольшее, что вы можете утверждать, так это то, что если бы с внедрением новой системы переменные издержки были точно равны \$6,27, то вы неверно определяли бы уровень значимости только в 5% случаев.

Пример. *Можно ли увеличить стоимость фирмы, изменив ее название?*

Когда крупная фирма меняет название, это является важным событием. Финансовые средства на рекламу в связи с изменением названия и созданием нового имиджа могут быть огромны. Почему фирмы идут на это? В соответствии с теорией финансового планирования фирмам следует предпринимать только такие проекты, которые увеличивает стоимость фирмы для ее собственников, для акционеров. Если фирма посчитала разумным потратить такие ресурсы на изменение своего названия, то должен наблюдаться рост стоимости фирмы, которая измеряется ценой ее акций.

Анализ изменения стоимости фирмы со времени объявления о перемене ее названия можно выполнить, используя одностороннюю статистическую проверку гипотезы, чтобы определить, действительно ли цена акций пошла вверх. Одна из трудностей измерений такого вида реакции рыночной цены состоит в том, что фондовый рынок подвержен влиянию различных сил, и оценивать влияние изменения названия следует с учетом всего фондового рынка. Цену акции следует сравнить с той, что могла бы быть в то время. Поэтому если фондовый рынок переживает подъем, то прежде чем принять решение о том, что объявление изменения названия эффективно, необходимо показать, что стоимость акций фирмы возросла на *значимо больший процент*, чем можно было бы ожидать.

Этот вид *анализа события*, учитывающий влияние широкомасштабных рыночных сил, включает вычисление *сверхдохода*, который представляет собой ставку дохода, получаемую инвестором-держателем акций фирмы за вычетом ставки дохода, которую можно было ожидать от инвестиций с аналогичным риском (но без изменения названия фирмы) за тот же период времени.

Таким образом, положительное значение сверхдохода будет свидетельствовать о том, что изменение названия фирмы вызвало рост цены акции в большей степени, чем можно было бы ожидать без такого изменения. Здесь возможны два случая. Фондовый рынок в целом переживает подъем, а акции фирмы еще больше идут вверх. Или на фондовом рынке спад, но акций фирмы он коснулся в меньшей степени, чем можно было бы ожидать, рассматривая рынок в целом.

Рассмотрим исследование, анализирующее 58 корпораций, изменивших свое название с 1981 по 1985 гг.[15] Авторы исследования сформулировали свой подход следующим образом.

Чтобы проверить, действительно ли сверхдоход, обусловленный изменением названия фирмы, отличен от нуля, в качестве статистики для проверки используется отношение среднего значения сверхдохода... к его стандартному отклонению... . Эта статистика подчиняется нормальному распределению при достаточно большом размере выборки n .

Для проверки гипотезы относительно равенства среднего выборки заданному значению $\mu_0 = 0$, т.е. предположения о том, что сверхдохода, обусловленного изменением названия фирмы, нет, использовали метод t-статистики. Упоминаемое в описании метода "стандартное отклонение" представляет собой стандартную ошибку этой оцененной величины. Поскольку размер выборки достаточно большой, t-значение (для бесконечного n) равно соответствующему значению для нормального распределения.

Результаты исследования были представлены таким образом.

Среднее значение сверхдохода равно 0,61% при соответствующем значении t-статистики, равном 2,15. Таким образом, если нулевая гипотеза предполагает, что значения разности между ожидаемым и реальным доходами взяты из совокупности с неположительным (отрицательным или равным нулю) значением среднего, то одностороннюю нулевую гипотезу можно отклонить.

В связи с тем, что нулевая гипотеза отклонена, принимают альтернативную гипотезу. Показано, что в результате изменения названия фирмы цена акции значимо возросла. Означает ли это, что нужно торопиться немедленно менять название фирмы? Совсем необязательно. Основные выводы данного исследования формулируются следующим образом.

Исследование показало, что для большинства фирм изменение названия связано с улучшением результата их деятельности и наибольшее улучшение чаще всего наблюдается в фирмах, выпускающих промышленные товары, чьи результаты в период, предшествующий изменению, были относительно невысоки. ...Результаты исследования, однако, не свидетельствуют о том, что новое название само по себе увеличивает спрос на продукцию фирмы. Скорее, изменение названия служит сигналом к принятию других серьезных мер по улучшению результата экономической деятельности, а именно маркетинговых и организационных изменений.

Обратите внимание, что при значении t-статистики, равном 2,15, получен результат статистически значимый на уровне 5% (так как это значение превышает 1,645). Однако поскольку t-значение для одностороннего интервала на уровне 1% равно 2,326, то этот результат является значимым, но не высоко значимым.

[15] Horsky D. and Swyngedouw P. "Does It pay to Change Your Company's Name? A Stock Market Perspective", *Marketing Science* 6 (1987), p. 320–335.

10.5.4. Проверка того, принадлежит ли новое наблюдение той же генеральной совокупности

Теперь вы, вероятно, считаете, что построив доверительный интервал, можно выполнить проверку гипотезы. Это верно. Используя интервал предсказания, описанный в главе 9, для нового наблюдения (вместо среднего генеральной совокупности) теперь можно быстро проверить, взято ли новое наблюдение из той же генеральной совокупности, что и выборка, или нет. Нулевая гипотеза H_0 утверждает, что новое наблюдение принадлежит той же нормально распределенной генеральной совокупности, что и выборка, а альтернативная гипотеза H_1 утверждает, что это не так. Набор данных рассматривают как случайную выборку.

Теперь, когда вы умеете строить доверительные интервалы и знаете основы проверки статистических гипотез, такого рода проверка является достаточно простой. Исходя из данных выборки (но без учета нового наблюдения) и используя *стандартную ошибку предсказания* $S\sqrt{1+1/n}$, находят интервал предсказания (особый вид доверительного интервала), как это описано в главе 9. Затем берут новое наблюдение. Если новое наблюдение *не* попадает в доверительный интервал, то делают вывод, что новое наблюдение значимо отличается от других.

При проверке методом t-статистики значение t-статистики просто вычисляют по формуле

$$t_{\text{статистика}} = \frac{X_{\text{новое}} - \bar{X}}{S\sqrt{1+1/n}},$$

где в числителе записана разность значений нового наблюдения и среднего выборки, а в знаменателе — стандартная ошибка предсказания. Затем, как и ранее, сравнивают значение t-статистики с критическим значением из t-таблицы (для $n-1$ степеней свободы).

Если вам необходим односторонний (направленный) тест, чтобы утверждать, что новое наблюдение либо значимо больше, либо значимо меньше среднего значения остальных наблюдений, то просто найдите соответствующий односторонний интервал предсказания или сравните значение t-статистики с t-значением для одностороннего интервала.

Пример. *Находится ли данная система под контролем*

Вы схватились за голову. Обычно художественные изделия из литого фарфора, которые изготавливают на данном станке, весят около 30 фунтов каждое. Конечно, есть некоторые отклонения, не все изделия весят *точно* 30 фунтов, допускается, что эти однотипные изделия не абсолютно идентичны. Но это же просто позор! Обнаружено изделие весом 38,31 фунта, что намного выше ожидаемого. Вас беспокоит, не вышел ли процесс производства *из-под контроля*, или это только случайность, которую можно ожидать время от времени. Вы бы предпочли не регулировать оборудование, поскольку это связано с отключением сборочной линии и поиском дефекта, но если сборочная линия действительно вышла из-под контроля, то чем быстрее вы устраните причину, тем лучше.

Нулевая гипотеза предполагает, что система все еще управляема, т.е. что отклонение веса последнего изделия обусловлено обычной случайностью. Альтернативная гипотеза предполагает, что система неуправляема и последнее изделие значимо отличается от остальных. Ниже приведена информация относительно последнего изделия и выборки из обычной продукции.

Размер выборки, n	19
Среднее выборки, $\bar{X}$	31,52 фунта
Стандартное отклонение, S	4,84 фунта
Новое наблюдение, $X_{\text{новое}}$	38,31 фунта

Стандартная ошибка предсказания определяется по формуле:

$$\text{Стандартная ошибка предсказания} = S\sqrt{1+\frac{1}{n}} = 4{,}84\sqrt{1+\frac{1}{19}} = 4{,}97.$$

В данном случае не совсем корректно использовать одностороннюю проверку, поскольку вас интересуют изделия, превышающие стандартный вес, и изделия с меньшим весом, и в обоих случаях система может рассматриваться как неуправляемая. Двусторонний интервал предсказания, построенный с учетом t-значения, равного 2,101, находится между значениями $31{,}52-2{,}101\times4{,}97=21{,}1$ и $31{,}52+2{,}101\times4{,}97=42{,}0$.

Мы на 95% уверены, что новое наблюдение, взятое из той же генеральной совокупности, что и выборка, будет находиться между 21,1 и 42 фунтами.

Новое наблюдение, 38,31 фунта, находится в интервале предсказания, а значит, и в пределах допустимых отклонений. Хотя значение веса ближе к верхней границе, оно *незначимо отличается* от остальных.

Значение t-статистики меньше (по абсолютной величине) критического значения, равного 2,101, что подтверждает выбранное решение принять нулевую гипотезу:

$$t_{\text{статистика}} = \frac{38{,}31-31{,}52}{4{,}965735} = 1{,}367.$$

Ретроспективно не следует удивляться весу изделия, равному 38,31 фунта. Поскольку стандартное отклонение выборки равно 4,84 фунта, то значения веса отдельных изделий будут довольно существенно отличаться от среднего значения веса. Вес рассматриваемого изделия не отклоняется даже на две стандартные ошибки от среднего и поэтому (даже в соответствии с этим приближенным правилом) находится в пределах допустимой 95% области. Конечно, такие рассуждения весьма приблизительны. Использование стандартной ошибки для предсказания дает возможность получить точный ответ, потому что в таком случае вы математически корректно учитываете и отклонения от среднего в вашей выборке и отклонение нового наблюдения.

10.6. Сравнение двух выборок

Чтобы проверить, значимо ли различаются с точки зрения среднего значения между собой две выборки, необходимо знать: (1) соответствующую стандартную ошибку, чтобы оценить среднюю разность; и (2) число степеней свободы. Затем задача, по существу, сводится к уже рассмотренным ранее: необходимо проверить гипотезу о равенстве наблюдаемой величины (наблюдаемой средней разности) известному заданному значению (нулю, который свидетельствует об отсутствии различия) с использованием соответствующей стандартной ошибки и критического значения из t-таблицы.

Можно заметить, что этот метод все время повторяется в статистике. При наличии оценки некоторой величины и ее стандартной ошибки легко можно построить доверительный интервал и выполнить проверку статистической гипотезы. Приложения могут усложняться (и становиться более интересными), но методы практически не меняются. Давайте обобщим этот метод на случай двух выборок.

t-тест для зависимых выборок

t-тест для двух зависимых выборок используют, чтобы проверить, различаются ли два столбца чисел с точки зрения среднего значения при условии, что *числа в двух столбцах образуют пары*. Такая ситуация возникает, например, в исследований типа "до/после", где рассматривается результат измерения некоторой величины (оценки в результате тестирования или рейтинга) для каждого объекта как до, так и после некоторого вмешательства (например, просмотр рекламы, проведение лечения, регулировка прибора и т.п.).

Фактически вам уже известно, как выполнить t-тест для зависимых выборок, поскольку его можно свести к известной нам проверке для одной выборки. Для этого нужно перейти к работе с *разностями* (например, из значения "после" вычесть значение "до"), вместо того, чтобы работать с каждой выборкой отдельно. Главное, чтобы четко было видно, как объединяются значения в двух выборках в пары. Иначе будет неясно, для каких пар вычислять разность.

Недостаточно иметь средние и стандартные отклонения для каждой из этих двух групп. В таком случае мы не учитываем информацию, привнесенную тем фактом, что наблюдения объединены в пары. Поэтому мы будем работать со средним и стандартным отклонением *разностей*.

t-тест для зависимых выборок может быть очень эффективен, даже если значения в группах сильно отличаются друг от друга. Поскольку эта проверка концентрируется на *изменениях*, она может игнорировать (потенциально сбивающую с толку) вариацию на уровне *отдельных* объектов. Например, отдельные личности могут быть совершенно разными, а анализируемые для них изменения — подобными (например, каждый получает прибавку в $100). Выявляя систематические изменения, t-тест для зависимых выборок не отвлекается на изменчивость признаков для отдельных объектов в выборках.

Корректное использование t-теста для зависимых выборок требует выполнения некоторых условий. Первое условие заключается в том, что изучаемые нами объекты представляют собой случайную выборку, извлеченную из изучаемой генеральной совокупности. Каждому объекту соответствуют два измерения признака. Далее, рассматривают набор данных, состоящий из разностей между этими двумя наборами измерений признака. Второе условие заключается в том, что эти разности распределены (по крайней мере приблизительно) нормально.

Пример. *Реакция на рекламу*

Необходимо определить эффективность рекламы в плане создания ею настроения расслабленности (релаксации). Выборку из 15 человек опросили до и после просмотра рекламного ролика. Вопросник включал много пунктов, в одном из которых респондентов просили описать свое состояние по шкале от 1 (напряженное состояние) до 5 (полностью расслабленное состояние). Результаты опроса приведены в табл. 10.6.1 (обратите внимание, что средняя оценка расслабленности увеличивается на 0,67, от 2,8 перед просмотром до 3,47 после просмотра).

Задача выглядит как сравнение двух выборок, но это не так. В действительности это задача, связанная с анализом изменений оценок расслабленности в одной выборке. Например, респондент №1 изменил оценку своей расслабленности с 3 на 2, т.е. для него изменение оценки расслабленности равно –1. (Обычно вычисляют разность между оценкой "после" и оценкой "до", и таким образом увеличение оценки представляют положительным числом, а уменьшение — отрицательным.) Вычислив разность для каждого респондента, мы перейдем к знакомой нам задаче для одной выборки (табл. 10.6.2).

Таблица 10.6.1. Оценки расслабленности

	До	После
Респондент №1	3	2
Респондент №2	2	2
	2	2
	4	5
	2	4
	2	1
	1	1
	3	5
	3	4
	2	4
	5	5
	2	3
	4	5
	3	5
Респондент №15	4	4
Размер выборки	15	15
Среднее	2,8000	3,4667
Стандартное отклонение	1,0823	1,5055

Таблица 10.6.2. Изменение оценки расслабленности

	После – До
Респондент №1	–1
Респондент №2	0
	0
	1
	2
	–1
	0
	2
	1
	2
	0
	1
	1
	2
Респондент №15	0
Размер выборки	15
Среднее	0,6667
Стандартное отклонение	1,0465

Нам известно, как решать такую задачу для одной выборки. Находим по таблице t-значение для двустороннего интервала (2,145) и, учитывая, что среднее выборки $\bar{X} = 0,6667$, а стандартная ошибка $S_{\bar{x}} = 0,2702$, делаем вывод.

Мы на 95% уверены, что среднее изменение оценки расслабленности для большей генеральной совокупности находится в пределах от 0,087 до 1,25.

Чему в данном случае равно заданное значение μ_0? Здесь $\mu_0 = 0$, потому что равное нулю значение изменения свидетельствует об *отсутствии влияния* просмотра рекламы на расслабленность в генеральной совокупности.

Проверка гипотезы состоит просто в том, чтобы определить, попадает заданное значение $\mu_0 = 0$ в пределы доверительного интервала или нет. В нашем случае не попадает, а следовательно, результат является значимым. Таким образом, если исходить из имеющихся данных, то "0" не является приемлемым значением для изменения расслабленности в генеральной совокупности.

Просмотр рекламы значимо увеличивает расслабленность ($p < 0,05$, двусторонняя проверка).

В данном случае необходима двусторонняя проверка, потому что нас также интересует, не вызвала ли реклама значимого *снижения* расслабленности. Определив значимость с помощью двустороннего теста, можно сделать и одностороннее заключение.

Для полноты картины приведем формулировки гипотез. Нулевая гипотеза $H_0 : \mu = 0$ утверждает, что среднее значение изменения между оценками расслабленности "до просмотра" и "после просмотра" в генеральной совокупности равно нулю, т.е. нет изменения средней оценки расслабленности. Альтернативная гипотеза $H_1 : \mu \neq 0$ утверждает, что есть изменение средней расслабленности между "до просмотра" и "после просмотра".

t-тест для независимых выборок

t-тест для двух независимых выборок используют, чтобы установить, существует ли различие средних для двух *независимых* столбцов чисел. Значения в этих двух столбцах нельзя естественным образом объединить в пары. Например, имеются данные о фирмах в двух промышленных группах, или необходимо сравнить выборки, взятые из продукции двух различных производственных линий. В таких случаях нельзя сводить данные в один столбец чисел — необходимо работать с двумя выборками.

Как только определена соответствующая стандартная ошибка, остальное выполнить уже легко. У вас есть оценка (разность между средними двух выборок), ее "собственная" стандартная ошибка и соответствующее число степеней свободы. Остается только построить доверительный интервал и проверить гипотезу.

У нас есть две выборки, выборка 1 и выборка 2. Основные статистики для обоих выборок обозначаются обычным способом, как показано в табл. 10.6.3.

Рассмотрим, что в этом случае появилось нового. Стандартная ошибка разности указывает на выборочную изменчивость *разности* между двумя выборочными средними. Есть две различные формулы: формула для большой выборки, которую используют, когда размер каждой из двух выборок не менее 30, и формула для малой выборки, которую используют в предположении, что обе генеральные совокупности имеют одинаковую изменчивость.[16] Формула для большой выборки работает даже тогда, когда изменчивость у выборок разная, за счет непосредственного объединения двух стандартных ошибок $S_{\bar{X}_1}$ и $S_{\bar{X}_2}$. Чтобы оценить изменчивость генеральной совокупности (при допущении, что она одинакова для обеих генеральных совокупностей), формула для малой выборки включает взвешенное среднее выборочных стандартных отклонений. Стандартная ошибка для случая малой выборки имеет $n_1 + n_2 - 2$ степеней свободы: из объединенного размера двух выборок $n_1 + n_2$ дважды вычитают 1 (для каждой оценки выборочного среднего). Ниже приведены формулы вычисления стандартной ошибки для каждого случая.

Стандартная ошибка разности

Выборка большого размера ($n_1 \geq 30$ и $n_2 \geq 30$):

$$S_{(\bar{X}_2-\bar{X}_1)} = \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

$$S_{(p_1-p_2)} = \sqrt{S_{p_1}^2 + S_{p_2}^2}$$ (для двух выборок с биномиальным распределением)

Количество степеней свободы — бесконечное

Выборка малого размера (в предположении о равенстве изменчивости):

$$S_{(\bar{X}_2-\bar{X}_1)} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

Количество степеней свободы — $n_1 + n_2 - 2$.

[16] Существуют решения и для малых выборок с неодинаковой изменчивостью, но эти решения более сложные. Один из подходов описан в книге Snedecor G. W. and Cochran W. G. "Statistical Methods", 6th ed. (Ames: Iowa State University Press, 1976), p. 115.

Таблица 10.6.3. Система обозначений для двух выборок

	Выборка 1	Выборка 2
Размер выборки	n_1	n_2
Среднее	$\bar{X}_1$	$\bar{X}_2$
Стандартное отклонение	S_1	S_2
Стандартная ошибка	$S_{\bar{X}_1}$	$S_{\bar{X}_2}$
Среднее разности	$\bar{X}_2 - \bar{X}_1$	

Будьте внимательны, используя в каждой формуле соответствующий показатель изменчивости: либо выборочное стандартное отклонение, либо стандартную ошибку выборки. Формула для большой выборки демонстрирует, как можно использовать оба этих показателя. Если в случае выборки малого размера у вас есть стандартные ошибки, а не стандартные отклонения, преобразуйте их в соответствующие стандартные отклонения, умножив на корень квадратный из размера соответствующей выборки. Обратите внимание, что в обеих формулах значения стандартных отклонений перед их сложением возводятся в квадрат.[17]

В формуле вычисления стандартной ошибки для случая выборки большого размера оценки *дисперсий* оценок $\bar{X}_1$ и $\bar{X}_2$ складывают для получения оценки дисперсии разности. Извлекая из этой суммы квадратный корень, находят оценку стандартного отклонения разности, которая и дает стандартную ошибку разности.

В формуле вычисления стандартной ошибки для случая выборки малого размера первый сомножитель в подкоренном выражении представляет собой средневзвешенную сумму стандартных отклонений (взвешенную в соответствии с числом степеней свободы для каждого из них). Остальная часть формулы преобразует изменчивость отдельных *элементов* в изменчивость *средней разности* путем сложения обратных значений размеров выборок, делая дважды то, что нужно было бы сделать один раз для вычисления обычной стандартной ошибки.

Проверяется гипотеза $H_0 : \mu_1 = \mu_2$ против альтернативной гипотезы $H_1 : \mu_1 \neq \mu_2$. Можно также записать в эквивалентной форме: $H_0 : \mu_2 - \mu_1 = 0$ против $H_1 : \mu_2 - \mu_1 \neq 0$. Предварительные условия, выполнение которых необходимо для t-проверки двух независимых выборок, те же, что были рассмотрены ранее, с добавлением одного нового, но только для случая выборок малого размера. Во-первых, предполагается, что каждая выборка является случайной выборкой из своей генеральной совокупности. (Здесь имеются две генеральные совокупности и две независимые выборки, представляющих эти совокупности.) Во-вторых, предполагается, что каждое выборочное среднее распределено приблизительно нормально, как мы и требовали ранее. И, наконец, только для случая выборок малого размера предполагается, что в двух генеральных совокупностях *стандартные отклонения равны* между собой, $\sigma_1 = \sigma_2$. Иными словами, две гене-

[17] Таким образом, как и во многих других формулах, усредняются *дисперсии*. В связи с этим специалисты, занимающиеся теоретической статистикой, обращают прежде всего внимание на дисперсию. Однако, чтобы получить содержательную интерпретацию таких чисел в осмысленных единицах измерения, необходимо извлечь квадратный корень. Вот почему в этой книге мы в основном работаем со стандартным отклонением, а не с дисперсией. Обратите внимание, что дисперсия и стандартное отклонение несут одинаковую информацию, поскольку их легко преобразовывать друг в друга.

ральные совокупности отличаются (если отличаются) только своими средними значениями, но не изменчивостью отдельных элементов по отношению к среднему генеральной совокупности.

Пример. *Дискриминация по полу и заработная плата*

Вашей фирме предъявлено обвинение в дискриминации сотрудников по признаку пола, и вам поручено изучить документы, представленные другой стороной. Документы включают проверку статистической гипотезы относительно размера заработной платы мужчин и женщин, которая демонстрирует "высоко значимую разницу" в средних значениях размера заработной платы мужчин и женщин. В табл. 10.6.4 приведены результаты этой проверки.

Таблица 10.6.4. Размеры заработной платы мужчин и женщин (в долларах)

	Женщины		Мужчины
	21 100		38 700
	29 700		30 300
	26 200		32 800
	23 000		34 100
	25 800		30 700
	23 100		33 300
	21 900		34 000
	20 700		38 600
	26 900		36 900
	20 900		35 700
	24 700		26 200
	22 800		27 300
	28 100		32 100
	25 000		35 800
	27 100		26 100
			38 100
			25 500
			34 000
			37 400
			35 700
			35 700
			29 100
Размер выборки	15		22
Среднее	$24 467		$33 095
Стандартное отклонение	$2 806		$4 189
Стандартная ошибка	$ 724		$ 893
Среднее значение разности		$ 8 628	

В этом отделе работают 15 женщин и 22 мужчины, средний годовой размер зарплаты составляет для женщин $24476 и для мужчин — $33095. В среднем мужчины зарабатывают на $8628 больше женщин. Таковы факты. Однако проблема заключается в том, является ли эта разница обычным случайным отклонением или нет. По существу, не имеет значения, как разделить эту группу из 37 человек на две группы по 15 и 22 человека, чтобы найти разницу средних размеров заработной платы. Вопрос в том, может ли такая большая разница в размерах заработной платы быть результатом лишь *случайного* распределения размеров заработной платы между мужчинами и женщинами, или необходимо другое объяснение этого очевидного неравенства.

Стандартные отклонения ($2806 для женщин и $4189 для мужчин) показывают, на какую приблизительно сумму отличаются размеры заработной платы отдельных людей в каждой группе. Большее колебание в размере заработной платы наблюдается среди мужчин, но его недостаточно, чтобы мы отказались от выполнения t-теста для двух независимых выборок.

Стандартные ошибки ($724 для женщин и $893 для мужчин) показывают, насколько сильно отличается среднее значение заработной платы в каждой из групп от среднего значения заработной платы в соответствующих идеализированных генеральных совокупностях. Например, если рассматривать группу из 15 человек как случайную выборку, извлеченную из идеализированной генеральной совокупности женщин, находящихся в аналогичных условиях, то среднее значение заработной платы для женщин $24 467 (в случайной выборке, потому что проанализировали размеры заработной платы только 15 человек), будет приблизительно на $724 отличаться от среднего в идеализированной генеральной совокупности.

Очевидно, что мы имеем дело с двумя *независимыми* выборками. Хотя можно было бы отнять значение размера заработной платы Мэри от значения размера заработной платы Джима, систематического обоснованного способа построения такого рода пар нет, поскольку реально мы имеем дело с двумя отдельными независимыми группами.

Оценим среднее значение разности, равное $8 628, чтобы решить, является ли она случайной. Для этого необходимо знать стандартную ошибку и число степеней свободы. Ниже представлен расчет, выполненный по формуле для случая выборки малого размера.

$$S_{\bar{x}_2-\bar{x}_1}=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=$$

$$=\sqrt{\frac{(15-1)2806^2+(22-1)4189^2}{15+22-2}\left(\frac{1}{15}+\frac{1}{22}\right)}=$$

$$=\sqrt{\frac{(14)7873636+(21)17547721}{35}(0{,}066667+0{,}045455)}=$$

$$=\sqrt{13678087\times 0{,}112121}=$$

$$=\sqrt{1533603}=$$

$$=1238.$$

Число степеней свободы равно $n_1+n_2-2=15+22-2=35$.

t-значение для 99,9% доверительного интервала равно 3,591. Помните, что t-значение находят по таблице с учетом числа *степеней свободы*, так как мы имеем дело с двумя выборками. Доверительный интервал находится между $8628-3591\times 1{,}238$ и $8628+3591\times 1{,}238$.

> Мы на 99,9% уверены, что разность средних значений размеров заработной платы в генеральных совокупностях находится между $4 182 и $13 074.

Этот доверительный интервал *не* включает заданное значение, равное 0, означающее отсутствие различия между средними значениями размеров заработной платы мужчин и женщин. Таким образом, мы принимаем следующее решение относительно результата проведенной проверки гипотез.

> Разница между средними размерами заработной платы мужчин и женщин является очень высоко значимой ($p<0{,}001$).

Этот результат подтверждает и тот факт, что значение t-статистики 8628/1238 = 6,97 намного больше, чем критическое для уровня значимости 0,001 табличное t-значение 3,591.

Какой вывод можно сделать из этого? Во-первых, распределение размеров заработной платы между мужчинами и женщинами не является случайным. Его можно было быть считать случайным, но только если допустить, что произошло очень редкое, встречающееся 1 раз на 1000 случаев, событие (так как именно такой смысл имеет уровень значимости 0,001). Во-вторых, если распределение размеров заработной платы не случайно, то должно быть некоторое объяснение. И здесь каждый человек может выдвинуть свою причину, думая что она полностью доказана результатами этого теста. Однако одно дело сказать, что причина есть, а другое — указать, что это за причина. Статистика исключает случайность как приемлемую возможность. И это все. Если вам хочется выдвинуть причину наблюдаемой разности размеров заработной платы, это ваше право, но это уже вне сферы статистики. Предложив основание для объяснения, статистика "уходит" на задний план (как с закатом солнца уезжает одинокий Рейнджер).

Итак, что может быть причиной разницы в размерах заработной платы? Одно из объяснений заключается в том, что руководство из консервативных и эгоистических соображений противозаконными способами умышленно решило платить отдельным служащим меньше только потому, что они женщины, руководствуясь в этом своем решении только полом служащего. Но это не единственно возможное объяснение. Разница в размерах заработной платы может быть обусловлена другими факторами, которые (1) влияют на размер заработной платы и (2) связаны с полом человека. В свою защиту фирма может заявить, что она устанавливает заработную плату только на основании *образования* и *опыта работы* и не ее вина, что при подборе кадров среди претендентов мужчин более образованных и опытных больше, чем среди претендентов женщин. Этот аргумент перемещает обвинение с фирмы на общество в целом.

Это сложная проблема. К счастью (для автора), мы не будем пытаться решить данный вопрос в этой книге. В принципе, это не вопрос статистики, и его нужно решать с привлечением экспертов из других областей. В книге мы еще вернемся к этому вопросу в главе о множественной регрессии, продолжив попытки понять взаимодействие таких факторов, как пол, размер заработной платы, образование и стаж работы.

Статистика очень полезна для получения точных ответов в условиях неопределенности, но эти ответы ограничены, и может понадобиться много работы и размышлений, прежде чем будет получен окончательный результат.

Пример. *Производительность конкурирующих отделов*

Вы дружески соревнуетесь с менеджером другого отдела, пытаясь выяснить, в чьем отделе производительность труда сотрудников выше. Фактически это соревнование не совсем дружеское, потому что у вас один начальник и он принимает решение о распределении ресурсов на основе достигнутых результатов. Вам хочется иметь не просто высокую, а *значимо* более высокую производительность, чтобы не возникал вопрос, чьи сотрудники успевают сделать больше.[18] Ниже приведены данные, характеризующие производительность труда служащих двух подразделений.

	Ваш отдел	Конкурирующий отдел
Размер выборки	53	61
Среднее	88,23	83,70
Стандартное отклонение	11,47	9,21
Стандартная ошибка	1,58	1,18
Среднее значение разности	4,53	

[18] На практике даже после проверки гипотезы остаются *некоторые* вопросы, потому что всегда есть возможность совершить ошибку (I или II рода). Вы же надеетесь показать, что высочайшая средняя производительность ваших сотрудников не может быть объяснена лишь случайностью.

Оценим среднее разности оценок продуктивности, чтобы посмотреть, не обусловлено ли это значение только случайностью. Для этого нам необходима стандартная ошибка этой разности. Произведем расчет стандартной ошибки по формуле для выборки большого размера, так как размеры обеих выборок больше 30.

$$S_{(\bar{X}_2-\bar{X}_1)} = \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2} =$$
$$= \sqrt{1,58^2 + 1,18^2} =$$
$$= \sqrt{2,48 + 1,39} =$$
$$= \sqrt{3,87} =$$
$$= 1,97.$$

По таблице для 95% доверительного интервала находим t-значение, которое равно 1,960. Доверительный интервал находится между $4,53 - 1,960 \times 1,97$ и $4,53 + 1,960 \times 1,97$.

Мы на 95% уверены, что среднее значение разности производительности в генеральной совокупности находится между 0,67 и 8,39.

Этот доверительный интервал не включает значение 0, которое свидетельствует об отсутствии разницы между средними оценками производительности двух отделов в идеализированной совокупности. Таким образом, исходя из результата проверки гипотезы вы принимаете следующее решение.

Среднее значение разности между производительностью служащих вашего отдела и производительностью служащих конкурирующего отдела является статистически значимым.

Метод t-статистики, конечно же, дает тот же ответ. Здесь значение t-статистики 4,53/1,97 = 2,30 превышает критическое значение 1,960.

Можно также сделать одностороннее заключение о значимости на основе результата двусторонней проверки гипотезы. Поскольку производительность служащих вашего подразделения выше, то, говоря языком статистики, производительность труда в вашем отделе *значимо выше*, чем у ваших соперников. Поздравляю!

10.7. Дополнительный материал

Резюме

Проверка статистических гипотез использует данные для того, чтобы принять решение о выборе между двумя возможностями (которые называют *гипотезами*). Эту процедуру часто применяют для того, чтобы отличить структуру от простой случайности, и с этой точки зрения она дает возможность получить полезные исходные данные для принятия решения. Гипотеза представляет собой утверждение о генеральной совокупности, которое может быть либо верным, либо неверным. Данные помогают решить, какую из двух гипотез принять в качестве истинной. Нулевая гипотеза, которую обозначают H_0, представляет собой *опровергаемое*, часто очень конкретное утверждение, как, например, утверждение о чистой случайности. Исследовательская гипотеза, или альтернативная гипотеза, H_1, является целью доказательства, и для принятия ее в качестве истинной требуется убедительное доказательство против нулевой гипотезы H_0. Принятие нулевой гипотезы представляет собой слабый вывод, а отклонение нулевой гипотезы и принятие альтернативной представляет собой строгое заключение и значимый результат.

Чтобы выяснить, равно ли среднее генеральной совокупности μ заданному значению μ_0, проверяют нулевую гипотезу $H_0: \mu = \mu_0$ против альтернативной

гипотезы $H_1: \mu \neq \mu_0$. Заданное значение — это известное фиксированное число μ_0, которое получено не из выборочных данных. Такая проверка является двусторонней, поскольку альтернативная гипотеза позволяет, чтобы значение среднего для генеральной совокупности располагалось как справа, так и слева от заданного значения. Такую проверку гипотезы о среднем генеральной совокупности называют также t-тестом, или t-тестом Стьюдента. Проверка гипотезы заключается в выяснении того, дальше ли отстоит среднее значение выборки $\bar{X}$ от заданного значения μ_0, чем это могло бы быть вызвано случайностью при условии, что μ равно заданному значению μ_0. Таким образом, расстояние между $\bar{X}$ и μ_0 сравнивают со стандартной ошибкой $S_{\bar{X}}$, используя при этом t-таблицу. Процедуру проверки можно выполнить либо на основе двустороннего доверительного интервала (см. главу 9), либо с помощью t-статистики, которая вычисляется по формуле

$$t_{статистика} = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}.$$

Ниже описано выполнение двусторонней проверки с использованием доверительного интервала и с использованием t-статистики (оба подхода всегда дают один и тот же результат).

- Если заданное значение μ_0 находится в пределах двустороннего доверительного интервала, или (эквивалентное утверждение) $|t_{статистика}| < t_{табл}$, то нулевую гипотезу H_0 принимают как приемлемую возможность. Выборочное среднее $\bar{X}$ *незначимо* отличается от μ_0. Наблюдаемая разница между выборочным средним $\bar{X}$ и заданным значением μ_0 может быть обусловлена просто случайностью. Результат *не является статистически значимым.*
- Если заданное значение μ_0 не находится в пределах двустороннего доверительного интервала, или (эквивалентное утверждение) $|t_{статистика}| > t_{табл}$, то принимают альтернативную гипотезу H_1 и отвергают нулевую гипотезу H_0. Выборочное среднее $\bar{X}$ *значимо* отличается от μ_0. Наблюдаемая разница между выборочным средним $\bar{X}$ и заданным значением μ_0 не может быть обусловлена только случайностью. Результат *является статистически значимым.*

Выполняя проверку статистической гипотезы, вы, таким образом, принимаете нулевую гипотезу ($\mu = \mu_0$) всякий раз, когда μ_0 представляет собой приемлемо возможное значение для μ. Если нулевая гипотеза верна, вероятность того, что будет принято верное решение, равна доверительному уровню (95% или любому другому) того столбца, который вы использовали в t-таблице.

В табл. 10.7.1 содержатся основные величины, необходимые для проверки гипотез как о среднем нормально распределенной генеральной совокупности, так и о вероятности наступления события для биномиального распределения.

t-статистика представляет собой пример более общего понятия *тест-статистики*, наиболее полезной с точки зрения выбора одной из двух гипотез величины, которую вычисляют на основе имеющихся данных. Значение тест-статистики сравнивают с соответствующим критическим значением, которое находят по стандартной статистической таблице; например, t-значение из t-

Таблица 10.7.1. Проверка гипотез о среднем нормально распределенной генеральной совокупности и о вероятности наступления события для биномиального распределения

	Нормальное распределение	Биномиальное распределение
Среднее генеральной совокупности	μ	π
Заданное значение	μ_0	π_0
Нулевая гипотеза	$H_0: \mu = \mu_0$	$H_0: \pi = \pi_0$
Альтернативная гипотеза	$H_1: \mu \neq \mu_0$	$H_1: \pi \neq \pi_0$
Данные	$X_1, \ldots, X_n$	X событий из n испытаний
Оценка	$\bar{X}$	$p = \frac{X}{n}$
Стандартная ошибка	$S_{\bar{X}} = S/\sqrt{n}$	$S_p = \sqrt{p(1-p)/n}$
Доверительный интервал	от $\bar{X} - tS_{\bar{X}}$ до $\bar{X} + tS_{\bar{X}}$	от $p - tS_p$ до $p + tS_p$
t-статистика	$t = (\bar{X} - \mu_0)/S_{\bar{X}}$	$t = (p - \pi_0)/S_p$

таблицы является критическим t-значением. Полезное эмпирическое правило заключается в том, что если абсолютное значение t-статистики больше числа 2, то нулевую гипотезу отклоняют, в противном случае — принимают.

В зависимости от того, какая из гипотез в действительности является истинной, можно совершить два типа ошибок. Ошибку I рода допускают, отвергая верную нулевую гипотезу и объявляя результат проверки статистически значимым. Вероятность совершения ошибки I рода (при верной нулевой гипотезе) определяется выбором соответствующего значения в t-таблице, обычно это уровень 5%. Ошибку II рода совершают, принимая нулевую гипотезу и объявляя результат статистически *незначимым*, в то время как истинной является альтернативная гипотеза. Вероятностью совершения ошибки II рода (при верной альтернативной гипотезе) управлять нелегко, но она может находиться (в зависимости от истинного значения μ) в пределах от 0 до уровня доверия теста (например, 95%). Обратите внимание, что каждый тип ошибки основан на предположении об истинности одной из гипотез. Поскольку каждая из гипотез либо верна, либо неверна, в зависимости от генеральной совокупности (но *не* от данных), мы не рассматриваем понятие вероятности справедливости гипотезы.

Предварительные условия, необходимые для проверки гипотезы, следующие: (1) набор данных является случайной выборкой из рассматриваемой генеральной совокупности, (2) либо измеряемые величины являются приблизительно нормально распределенными, либо размер выборки достаточно велик для того, чтобы в соответствии с центральной предельной теоремой среднее значение выборки было приблизительно нормально распределено.

Уровень проверки, или уровень значимости, представляет собой вероятность принять альтернативную гипотезу, когда правильной является нулевая гипотеза

(т.е. совершить ошибку I рода). Обычно этот уровень устанавливают равным 5%, но его можно установить равным 1% или 0,1% (или даже 10% для некоторых задач), выбрав соответствующий столбец t-таблицы. p-значение показывает, насколько необычным является получение имеющихся данных при условии справедливости нулевой гипотезы. Малые p-значения свидетельствуют о малой вероятности такого события и приводят к отклонению H_0. Обычно H_0 отвергают, когда p-значение меньше 0,05. Результат проверки называют статистически значимым ($p < 0,05$), если он значим на уровне 5%. Используют также термины *высоко значимый* ($p < 0,01$), *очень высоко значимый* ($p < 0,001$) и *незначимый* ($p > 0,05$).

Односторонний t-тест соответствует нулевой гипотезе, утверждающей, что значение μ находится по одну сторону от μ_0, и альтернативной гипотезе, утверждающей, что значение μ находится по другую сторону от μ_0. Чтобы использовать односторонний (направленный) тест, следует быть уверенным, что *независимо от поведения данных*, вы будете продолжать использовать односторонний тест на той же стороне ("больше, чем" или "меньше, чем"). Если существуют сомнения, следует использовать двусторонний тест; если результат двустороннего теста окажется значимым, затем на его основе можно сделать *односторонний* вывод. Проверку гипотезы можно выполнить путем построения соответствующего одностороннего доверительного интервала (в соответствии с утверждением альтернативной гипотезы) или с помощью t-статистики. Значимый результат (принятие альтернативной гипотезы) будет иметь место, когда значение заданной величины μ_0 *не* попадет в доверительный интервал. Это происходит в том случае, когда $\bar{X}$ находится на предполагаемой альтернативной гипотезой стороне от μ_0 и значение t-статистики по абсолютной величине больше табличного t-значения. Результат является значимым, если $t_{статистика} > t_{табл}$ (когда проверяется $H_1 : \mu > \mu_0$) или $t_{статистика} < -t_{табл}$ (когда проверяется $H_1 : \mu < \mu_0$).

Для одностороннего t-теста, проверяющего, *больше* ли μ, чем μ_0, гипотезу формулируются так: $H_0 : \mu \le \mu_0$ и $H_1 : \mu > \mu_0$. Доверительный интервал включает все значения, которые по крайней мере не меньше, чем $\bar{X} - t_{табл} S_{\bar{X}}$.

- Если μ_0 находится внутри доверительного интервала, или (эквивалентное утверждение) $t_{статистика} \le t_{табл}$, то принимают нулевую гипотезу H_0 как приемлемую возможность. Среднее выборки $\bar{X}$ *незначимо больше* μ_0. Если $\bar{X}$ больше, чем μ_0, наблюдаемую разность можно объяснить только случайностью. Результат *не является статистически значимым*.
- Если μ_0 *не* находится внутри доверительного интервала, или (эквивалентное утверждение) $t_{статистика} > t_{табл}$, то принимают альтернативную гипотезу H_1, а нулевую гипотезу H_0 отвергают. Среднее выборки $\bar{X}$ *значимо больше* μ_0. Наблюдаемую разность *нельзя* объяснить лишь случайностью. Результат *является статистически значимым*.

Для одностороннего t-теста, проверяющего, *меньше* ли μ, чем μ_0, гипотезы формулируются таким образом: $H_0 : \mu \ge \mu_0$ и $H_1 : \mu < \mu_0$. Доверительный интервал включает все значения, которые *не больше*, чем $\bar{X} + t_{табл} S_{\bar{X}}$.

- Если μ_0 находится внутри доверительного интервала, или (эквивалентное утверждение) $t_{статистика} \ge -t_{табл}$, то принимают нулевую гипотезу H_0, как приемлемую возможность. Среднее выборки $\bar{X}$ *незначимо меньше* μ_0. Если $\bar{X}$ меньше, чем μ_0, наблюдаемую разность можно объяснить только случайностью. Результат *не является статистически значимым*.
- Если μ_0 *не* находится внутри доверительного интервала, или (эквивалентное утверждение) $t_{статистика} < -t_{табл}$, то принимают альтернативную гипотезу H_1, а нулевую гипотезу H_0 отвергают. Среднее выборки $\bar{X}$ *значимо меньше* μ_0. Наблюдаемую разность *нельзя* объяснить лишь случайностью. Результат *является статистически значимым*.

Имея оценку (как, например, $\bar{X}$), соответствующую стандартную ошибку (как, например, $S_{\bar{X}}$) и критическое значение из соответствующей таблицы (как, например, t-таблицы), можно построить одно- или двусторонний интервал (на различных уровнях доверительности) и выполнить одно- или двустороннюю проверку гипотезы (на различных уровнях значимости).

При проверке, принадлежит ли новое наблюдение той генеральной совокупности, из которой взята выборка, нулевая гипотеза утверждает, что принадлежит, а альтернативная гипотеза утверждает противоположное. Используя стандартную ошибку предсказания $S\sqrt{1+1/n}$, строят интервал предсказания, и, если новое наблюдение попадает в этот интервал, принимают нулевую гипотезу, в противном случае принимают альтернативную гипотезу и делают вывод о значимости. Можно также, используя приведенную ниже формулу, вычислить t-статистику и сравнить полученное значение с табличным t-значением.

Для проверки нового наблюдения:

$$t_{статистика} = \frac{X_{новое} - \bar{X}}{S\sqrt{1+1/n}}$$

Независимо от используемого метода (доверительные интервалы или t-статистика) вам доступны все те же уровни значимости, p-значения и одно- либо двусторонние тесты, что и ранее.

t-тест для двух зависимых выборок используют, чтобы проверить, равны ли средние генеральных совокупностей для таких двух выборок, у которых есть естественная связь между парами отдельных значений, — например, измерения "до" и "после" некоторого события, выполненные для одних и тех же людей. Работая с разностями значений ("после" минус "до"), проблему сводят к знакомому t-тесту для одной выборки, используя в качестве заданного значения $\mu_0 = 0$, что соответствует нулевой гипотезе об отсутствии различия между средними значениями.

t-тест для двух независимых выборок используют, чтобы проверить, равны ли средние генеральных совокупностей на основе двух таких выборок, у которых *нет* естественной связи между парами отдельных значений, т.е. речь идет о двух независимых выборках из двух разных генеральных совокупностей. При двусторонней проверке нулевая гипотеза утверждает, что разность средних равна 0. Чтобы построить доверительный интервал для разности средних и выполнить проверку гипотезы, необходимо вычислить стандартную ошибку разности (которая дает оценку стандартного отклонения выборочного среднего разности) и число степеней свободы для нее.

Для выборок большого размера ($n_1 \geq 30$ и $n_2 \geq 30$):

$$S_{(\bar{X}_2-\bar{X}_1)} = \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}};$$

$$S_{(p_1-p_2)} = \sqrt{S_{p_1}^2 + S_{p_2}^2} \quad \text{(для двух биномиальных распределений);}$$

число степеней свободы = бесконечное.

Для выборки малого размера (в предположении о равенстве изменчивости в двух генеральных совокупностях):

$$S_{(\bar{X}_2-\bar{X}_1)} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

число степеней свободы равно $n_1 + n_2 - 2$.

Исходя из среднего значения разности, стандартной ошибки разности, числа степеней свободы и заданного значения (0) строят доверительный интервал и выполняют проверку гипотезы обычным способом. Обратите внимание, что в случае выборки малого размера к обычным требованиям случайности выборки и нормальности распределения добавляется также требование равенства изменчивости в двух генеральных совокупностях ($\sigma_1 = \sigma$).

Основные термины

- Проверка статистической гипотезы (hypothesis testing), 446
- Гипотеза (hypothesis), 446
- Нулевая гипотеза (null hypothesis), 446
- Исследовательская гипотеза (research hypothesis), или альтернативная гипотеза (alternative hypothesis), 447
- Заданное значение (reference value), 450
- Двусторонняя проверка (two-sided test), 450
- t-тест (*t*-test), или тест Стьюдента (Student's test), 459
- Тест-статистика (test statistic), 459
- Критическое значение (critical value), 459
- Критическое t-значение (critical *t* value), 459
- t-статистика (*t* statistic), 459
- Ошибка I рода (type I error), 462
- Ошибка II рода (type II error), 463
- Предварительные условия проверки гипотез (assumptions for hypothesis testing), 464
- Статистически значимый (statistically significant), 465
- Уровень проверки (test level), или уровень значимости (significance level), 466
- Доверительная вероятность, или ρ -значение (ρ -value), 466

- Односторонний t-тест (one-sided *t* test), 469
- t-тест для двух зависимых выборок (paired *t* test), 480
- t-тест для двух независимых выборок (unpaired *t* test), 482
- Стандартная ошибка разности (standard error of the difference), 482

Контрольные вопросы

1. а) В чем заключается цель проверки статистических гипотез?

   б) В чем различие между результатом проверки статистической гипотезы и утверждением о доверительном интервале?
2. а) Что такое статистическая гипотеза? В частности, это утверждение о генеральной совокупности или о выборке?

   б) В чем отличие роли нулевой гипотезы от роли альтернативной (исследовательской) гипотезы? Какая из них, как правило, включает утверждение о чистой случайности? Какая из них требует доказательств? Какая из них имеет преимущества при наличии сомнений?

   в) Предположим, вы приняли решение в пользу нулевой гипотезы. Это сильное или слабое заключение?

   г) Предположим, вы приняли решение в пользу альтернативной гипотезы. Это сильное или слабое заключение?

   д) Прокомментируйте утверждение: "Нулевая гипотеза никогда не может быть опровергнута".
3. а) Кратко опишите этапы выполнения двусторонней проверки гипотезы о среднем генеральной совокупности с использованием доверительного интервала.

   б) Кратко опишите этапы выполнения двусторонней проверки гипотезы о среднем генеральной совокупности с использованием t-статистики.
4. а) Что такое t-тест Стьюдента?

   б) Кто такой Стьюдент? Какие результаты его работы вам известны?
5. а) Что такое заданное значение? Берут ли это значение из данных выборки? Известно это значение или нет?

   б) Что такое t-статистика? Зависит ли она от заданного значения?

   в) Изменяется ли доверительный интервал в зависимости от заданного значения?
6. а) Что такое в общих чертах тест-статистика?

   б) Какую тест-статистику следует использовать для двустороннего t-теста?

   в) Что такое в общих чертах критическое значение?

   г) Какое критическое значение следует использовать для двустороннего t-теста?
7. а) Какие предварительные условия должны быть выполнены для корректного выполнения двустороннего t-теста?

б) Рассмотрим каждое из этих условий отдельно. Что произойдет, если условие не выполняется? Что можно (если можно) предпринять для решения этой проблемы?

8. а) Что такое ошибка I рода? Можно ли ею управлять? Почему?

   б) Что такое ошибка II рода? Можно ли ею управлять? Почему?

   в) В каком случае (если это вообще возможно) можно сказать, что "нулевая гипотеза истинна с вероятностью 0,95"?

   г) Что вы можете сказать о корректных решениях по принятию верной нулевой гипотезы на протяжении всей вашей жизни?

9. Какое утверждение о ρ-значении ассоциируется с каждым из следующих результатов проверки статистической гипотезы?

   а) Незначимый.

   б) Значимый.

   в) Высоко значимый.

   г) Очень высоко значимый.

10. а) Что такое односторонний тест?

    б) Сформулируйте гипотезы для одностороннего теста.

    в) В каком случае допускается выполнение одностороннего теста? Что вы предпримите в случае сомнений?

    г) Если вы выполняете односторонний тест, когда его выполнение недопустимо, то что может произойти в наихудшем случае?

    д) При каких условиях допускается сделать одностороннее утверждение, основанное на двустороннем тесте?

11. а) Как выполняется односторонний тест на основе доверительного интервала?

    б) Как выполняется односторонний тест с помощью t-статистики?

12. Предположим, что имеется оценка некоторого параметра и вы хотите проверить, равно ли значение этого параметра в генеральной совокупности нулю. Что еще необходимо для этого знать?

13. Какую стандартную ошибку следует использовать для проверки того, взято ли новое наблюдение из той же совокупности, что и выборка? (Дайте название и укажите формулу.)

14. а) Что такое t-тест для двух зависимых выборок?

    б) Сформулируйте две гипотезы, входящие в t-тест для двух зависимых выборок.

    в) Как вы понимаете тот факт, что две выборки состоят из пар связанных между собой значений? Приведите конкретный пример.

    г) Чем t-тест для пары зависимых выборок похож и чем отличается от обычного t-теста для одной выборки?

15. а) Что такое t-тест для двух независимых выборок?

    б) Сформулируйте две гипотезы, входящие в t-тест для двух независимых выборок.

в) Как вы понимаете требование "независимость двух выборок"?

г) Чем t-тест для двух независимых выборок похож или чем отличается от обычного t-теста для одной выборки?

д) В каких случаях уместно использовать какие из стандартных ошибок? (Ответьте устно и приведите формулу.)

е) Какое дополнительное условие должно выполняться, чтобы можно было использовать процедуру t-теста для двух независимых малых выборок? Что нужно делать, если это условие сильно нарушается?

16. а) Опишите общий процесс построения доверительных интервалов и выполнения проверки гипотез с использованием эмпирического правила, учитывая, что у вас есть оценка и ее стандартная ошибка.

    б) Если известно число степеней свободы и можно использовать t-таблицу, как можно сделать ответ более точным?

Задачи

1. Чтобы нацелить маркетинговую кампанию своего ресторана на людей необходимой возрастной группы, вы хотите выяснить, существует ли статистически значимая разница между средним возрастом ваших клиентов и средним возрастом всех жителей города, который составляет 43,1 года. В случайной выборке из 50 ваших клиентов средний возраст равен 33,6 года со стандартным отклонением 16,2 года.

   а) Сформулируйте словами и в математических обозначениях нулевую и исследовательскую (альтернативную) гипотезы для двустороннего теста.

   б) Выполните двусторонний тест на уровне значимости 5% и опишите полученный результат.

2. а) Выполните для предыдущей задачи двусторонний тест на уровне значимости 1% и опишите результат.

   б) Установите ρ-значение либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$.

3. Часть сборочной линии необходимо регулировать, если консистенция подаваемого пластика становится слишком вязкой или, наоборот, недостаточно вязкой по сравнению со значением вязкости 56,00, которое ваши инженеры рассматривают как приемлемое. Вы решаете производить настройку только тогда, когда есть уверенность, что система "вышла из под контроля", т.е. когда существует реальная необходимость для регулировки. Среднее значение вязкости в последних 13 измерениях составила 51,22 со стандартным отклонением 3,18.

   а) Сформулируйте словами и в математических обозначениях нулевую и исследовательскую (альтернативную) гипотезы для двусторонней проверки.

   б) Выполните двусторонний тест при уровне значимости 5% и опишите полученный результат.

в) Выполните двусторонний тест при уровне значимости 1% и опишите полученный результат.

г) Установите ρ-значение либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$.

4. а) Почему двусторонний тест является приемлемым для предыдущей задачи?

б) Сформулируйте, если это возможно, односторонний вывод на основе результата двустороннего теста на уровне значимости 5%.

5. Похоже, что часть вашей рекламы оказалась недейственной, так как люди ее просто проигнорировали. Вы договорились об исследовании оценки знания людьми вашей торговой марки до и после просмотра телевизионного шоу, в котором была показана ваша реклама. Вы хотите узнать, имеет ли реклама статистически значимый эффект в сравнении с нулевым значением, которое представляет полное отсутствие эффекта. Когда 200 человек были опрошены до и после просмотра этой рекламы, то оказалось, что их осведомленность о вашей торговой марке, оцениваемая по шкале от 1 до 5, в среднем возросла на 0,22 балла. Стандартное отклонение этого увеличения составило 1,39 балла.

а) Сформулируйте словами и в математических обозначениях нулевую и исследовательскую (альтернативную) гипотезы для двусторонней проверки.

б) Выполните двусторонний тест на уровне значимости 5% и опишите полученный результат.

в) Выполните двусторонний тест на уровне значимости 1% и опишите полученный результат.

г) Установите ρ-значение либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$.

6. а) Почему двусторонний тест является приемлемым для этой задачи?

б) Сформулируйте, если это возможно, односторонний вывод на основе результата двустороннего теста на уровне значимости 5%.

7. Из вашей базы данных, содержащей информацию о 13916 клиентах, случайным образом отобрано для опроса 725 человек. Из них 113 человек ответили, что они недовольны сервисом вашей компании.

а) Вычислите наилучшую оценку процента недовольных сервисом клиентов во всей базе данных.

б) Найдите стандартную ошибку вашей оценки процента недовольных сервисом клиентов.

в) Вычислите наилучшую оценку количества недовольных сервисом клиентов во всей базе данных.

г) Определите 95% доверительный интервал для процента неудовлетворенных потребителей.

д) Компания поставила цель снизить процент недовольных потребителей до 10% или менее. Как вы считаете, эта цель уже достигнута или сейчас процент недовольных все еще выше 10%? Обоснуйте свой ответ.

8. Уровень материальных запасов на вашей фабрике в прошлом году определяли 12 раз в случайно выбранное время. Результаты таковы: 313, 891, 153, 387, 584, 162, 742, 684, 277, 271, 285, 845.

а) Определите обычный уровень материальных запасов для всего года, используя стандартные статистические показатели.

б) Определите генеральную совокупность.

в) Определите 95% доверительный интервал для среднего в генеральной совокупности значения уровня материальных запасов.

г) Значимо ли отличается среднее проведенных измерений уровня материальных запасов от значения 500, которое руководство использует для управления бюджетом? Обоснуйте свой ответ.

9. Ваша пекарня изготавливает буханки хлеба, на этикетках которых указан вес "один фунт". Ниже приведены значения веса случайно отобранных буханок из сегодняшней продукции:

1,02; 0,97; 0,98; 1,10; 1,10; 1,02; 0,98; 1,03; 1,03; 1,05; 1,02; 1,06.

а) Определите 95% доверительный интервал для среднего значения веса всех буханок, изготовленных сегодня.

б) Какое число нужно использовать в качестве заданного опорного значения при проверке гипотезы о сравнении среднего значений веса сегодняшних буханок с весом, указанным на этикетке?

в) Сформулируйте гипотезы H_0 и H_1.

г) Выполните проверку гипотезы (двусторонняя, уровень 0,05) и поясните полученные результаты.

д) Какую ошибку, если таковая может быть, вы могли совершить?

10. Можно ли существенно экономить деньги, делая покупки в большом универсальном магазине? Используйте данные из задачи 22 главы 9 для того, чтобы разобраться в этом вопросе

11. На последнем совещании принято решение о запуске нового изделия в случае, если "заинтересованные потребители согласятся платить в среднем \$20 за изделие". Изучение ответов 315 случайно выбранных заинтересованных потребителей показало, что они готовы платить в среднем \$18,14 за данное изделие. Стандартное отклонение составило \$2,98.

а) Укажите, какое заданное опорное значение нужно использовать при проверке гипотезы о среднем для всех заинтересованных потребителей.

б) Сформулируйте словами и в математических обозначениях нулевую и альтернативную гипотезы для двусторонней проверки.

в) Выполните двусторонний тест для уровня значимости 5% и опишите результат.

г) Выполните двусторонний тест для уровня значимости 1% и опишите результат.

д) Установите ρ-значение либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$.

12. а) Почему односторонний тест может быть приемлемым для предыдущей задачи?

б) Сформулируйте словами и в математических обозначениях нулевую и исследовательскую (альтернативную) гипотезы для односторонней проверки.

в) Выполните односторонний тест для уровня значимости 5% и опишите полученный результат.

13. Последний опрос случайно отобранных 809 зарегистрированных избирателей показал, что 426 человек планирует отдать свой голос за вашего кандидата на предстоящих выборах.

а) Превышает ли наблюдаемый процент 50%?

б) Значимо ли наблюдаемый процент превышает 50%? Как вы это определили? Обоснуйте ваш ответ с помощью двустороннего теста.

14. Проверьте, может ли процент в генеральной совокупности быть равным 20% исходя из того, что в опросе 500 случайно отобранных клиентов 18,4% ответили, что им ваши изделия нравятся.

15. Принимая решение о запуске в производство нового изделия, вам необходимо знать, достаточно ли большой процент (не менее 10%) жителей города будет заинтересован в его покупке. Изделие будет запущено в производство только в том случае, если вы убедитесь в достаточном спросе на него. Опрос 400 случайно отобранных жителей города показал, что 13% желают попробовать предложенное вами новое изделие.

а) Почему в этом случае будет уместна односторонняя проверка?

б) Сформулируйте словами и в математических обозначениях нулевую и альтернативную гипотезы для односторонней проверки.

в) Выполните проверку на уровне значимости 5% и опишите полученный результат.

г) Выполните проверку на уровне значимости 1% и опишите полученный результат.

д) Установите ρ-значение либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$.

16. Вы рассматриваете новую систему поставок и хотите проверить, существенно ли отличается среднее время поставки в рамках новой системы по сравнению с действующей. Известно, что среднее время поставки при действующей системе составляет 2,38 дня. Проверка новой системы на основе 48 наблюдений показывает, что среднее время поставки составляет 1,91 дня со стандартным отклонением 0,43 дня.

а) Сформулируйте словами и в математических обозначениях нулевую и альтернативную гипотезы для двусторонней проверки.

б) Выполните двусторонний тест на уровне значимости 5% и опишите полученный результат.

в) Выполните двусторонний тест на уровне значимости 1% и опишите полученный результат.

г) Установите ρ-значение либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$.

д) Подведите итоги, написав краткую докладную записку вашему руководству.

17. Вы работаете в компании по изготовлению и продаже замороженных продуктов питания. Чистый вес одной упаковки должен быть 14,5 унции. Ниже приведена случайная выборка значений веса упаковок из продукции, изготовленной сегодня:

14,43; 14,37; 14,38; 14,29; 14,60; 14,45; 14,16; 14,52; 14,19; 14,04; 14,31.

Выборка также была взята и из вчерашней продукции. Среднее составило 14,46, стандартное отклонение — 0,31.

а) Оцените средний вес, который вы могли бы получить в том случае, если бы у вас была возможность взвесить все упаковки, изготовленные за сегодняшний день.

б) Для обычной отдельной упаковки, произведенной вчера, определите, насколько отличается ее действительный вес от вчерашнего среднего?

в) Определите 95% доверительный интервал для среднего веса всех упаковок, произведенных сегодня.

г) Определите гипотезу, с которой необходимо работать для проверки того, является ли корректным (т.е. таким, каким он должен быть) в среднем вес упаковок в сегодняшней продукции.

д) Значимо ли различие между заявленным весом упаковки и средним весом упаковок в сегодняшней продукции? Обоснуйте свой ответ.

18. Выпускаемая вашей фирмой игра имеет заявленную в прайс-листах цену $12,95, однако каждый магазин может устанавливать цену по своему усмотрению. Вы провели небольшое исследование и получили следующие значения цены вашего изделия в случайной выборке магазинов, торгующих вашей продукцией:

12,95; 9,95; 8,95; 12,95; 12,95; 9,95; 9,95; 9,98; 13,00; 9,95.

а) Оцените среднюю цену продаж, которую вы могли бы получить, если бы имели возможность узнать цену во всех магазинах, торгующих вашей продукцией.

б) Насколько приблизительно отличается цена продажи игры в типичном магазине от средней?

в) Определите 95% доверительный интервал для средней цены продаж во всех магазинах, торгующих вашим изделием.

г) Ваш отдел маркетинга считает, что игры продаются со средней скидкой в 12% от цены, заявленной в прайс-листах. С какой гипотезой необходимо работать для сравнения среднего значения цены продаж в генеральной совокупности с этим мнением?

д) Проверьте гипотезу из п. "г".

19. Было случайно отобрано определенное количество произведенных за эту неделю замороженных обедов. Отобранные образцы были вскрыты для из-

мерения калорийности каждой упаковки. Одна упаковка должна содержать 200 калорий. Ниже приведены данные о количестве калорий в отобранных упаковках обедов:

221, 198, 203, 223, 196, 202, 219, 189, 208, 215, 218, 207.

а) Оцените среднее количество калорий, которое вы могли бы получить, если бы у вас была возможность измерить калорийность всех упаковок обедов, изготовленных за эту неделю.

б) Насколько приблизительно отличается среднее количество калорий в одном обеде для выборки от среднего количества калорий для всех обедов, произведенных на этой неделе?

в) Определите 99% доверительный интервал для среднего количества калорий в обедах, произведенных за текущую неделю.

г) Значима ли разница между требуемым и измеренным количеством калорий? Обоснуйте свой ответ.

20. Проанализируйте стоимость (в тысячах долларов) подарков, возвращенных после праздников в каждый из ваших универмагов (табл. 10.7.2).

а) Вычислите стандартное отклонение.

б) Представьте интерпретацию стандартного отклонения как меру изменчивости стоимости подарков, возвращенных в различные универмаги.

в) Вычислите стандартную ошибку среднего и кратко поясните полученное значение.

г) Определите двусторонний 95% доверительный интервал для среднего значения подарков, возвращенных во все универмаги.

д) Городская ассоциация торговцев ожидает, что, как это обычно было раньше, в среднем на один универмаг объем возврата будет составлять $10 000. Проверьте, значимо ли отличается это среднее значение за год от ожидаемого.

21. Ниже приведены оценки удовлетворенности 12 случайно отобранных клиентов:

89, 98, 96, 65, 99, 81, 76, 51, 82, 90, 96, 76. Значимо ли отличается наблюдаемое среднее значение оценки от запланированной оценки 80 баллов? Обоснуйте свой ответ.

Таблица 10.7.2. Стоимость возвращенных подарков (в тысячах долларов)

Универмаг	Возвращено
A	13
B	8
C	36
D	18
E	6
F	21

22. В соответствии с нормой ваша фабрика может сбрасывать загрязняющие окружающую среду отходы производства в количестве, не превышающем 25 мг вредного вещества в неделю. Последняя выборка дала следующие результаты недельных загрязнений: 13, 12, 10, 8, 22, 14, 10, 15, 9, 10, 6 и 12 мг.

а) Уложились ли вы в норму? Поясните ваш ответ исходя из односторонней проверки гипотезы на уровне 5%.

б) Установите p-значение либо как $p > 0{,}05$, либо как $p < 0{,}05$, либо как $p < 0{,}01$, либо как $p < 0{,}001$. Является ли полученный результат высоко значимым?

в) Сформулируйте необходимые гипотезы, а также допущения, сделанные при этих проверках.

г) К большему или меньшему количеству результатов, связанных с "выходом за пределы нормы", приведет использование двустороннего теста вместо одностороннего при прочих равных условиях? Объясните почему.

23. Считается, что производственный процесс находится "под контролем", если в течение длительного времени средний вес производимых деталей составляет 0,20 кг, хотя вес отдельных деталей может и отличаться от этого среднего. Ниже приведена случайная выборка недавно изготовленных деталей:

0,253; 0,240; 0,247; 0,183; 0,247; 0,223; 0,252; 0,195; 0,235; 0,241; 0,251; 0,261; 0,194; 0,236; 0,256; 0,241.

Можно ли утверждать, что производственный процесс находится под контролем? Обоснуйте свой ответ.

24. Объем производимой продукции может изменяться и быть в каждый конкретный день либо высоким, либо низким. Если он высокий, то вы хотите знать почему, чтобы так же увеличивать объем продукции и в другие дни. Если же он низкий, то следует разобраться в причине и устранить ее. Похоже, что сегодняшний объем продукции ниже, чем обычно. Какую проверку, одностороннюю или двустороннюю, следует использовать, чтобы убедиться в этом? Почему?

25. Последний опрос 1235 случайно отобранных потенциальных избирателей показал, что ваш кандидат впереди с 52,1% голосов. Баллотируются всего два кандидата. Чтобы сделать окончательный вывод относительно большей группы *всех* потенциальных избирателей, используйте проверку гипотез.

а) Аккуратно сформулируйте гипотезы для двусторонней проверки.

б) Выполните проверку гипотезы на уровне 0,05 и предоставьте результаты.

в) Тщательно сформулируйте точное утверждение, которое подводит итог и поясняет полученный результат проверки.

г) Проделайте процедуры, указанные в пп. "б" и "в", приняв, что за вашего кандидата готовы отдать голоса не 52,1%, а 58,3% опрошенных избирателей.

д) Объясните, почему в данном случае не подходит односторонний тест, показав, что каждый из трех возможных результатов двустороннего теста может представлять интерес.

26. Рассмотренный в этой главе пример показал, что, по словам менеджеров, наличие у служащих акций компании оказывает значимое позитивное влияние на качество выпускаемой продукции. В рамках этого же исследования менеджеров попросили оценить влияние наличия акций у служащих компании на *расходы на рабочую силу в расчете на единицу продукции*.[19] Этот эффект, оцененный по шкале от −2 (большой негативный эффект) до 2 (большой позитивный эффект), составил 0,12, со стандартной ошибкой 0,11 на основе выборки, включающей 343 менеджера.

а) Определите 95% доверительный интервал и дайте интерпретацию его смысла. Следует помнить, что это мнения случайно отобранных менеджеров.

б) Существует ли, по мнению менеджеров, значимая связь между наличием акций у служащих компании и расходами на рабочую силу в расчете на единицу продукции? Почему?

в) Определите нулевую и альтернативную гипотезы.

г) Какая из гипотез была принята? Это слабое или сильное заключение?

д) Полностью ли была доказана принятая гипотеза? Если нет, то какого типа ошибка могла быть совершена?

27. Цель проведения вашей маркетинговой кампании — добиться, чтобы более 25% покупателей в супермаркетах узнавали вашу торговую марку. Последний опрос 150 случайно отобранных покупателей показал, что только 21,3% знают вашу торговую марку.

а) Необходимо найти аргументы, показывающие, что более 25% покупателей узнают вашу торговую марку. Определите необходимую для этого случая одностороннюю гипотезу и выполните проверку на уровне 0,05.

б) С другой стороны, возможно, интересно рассмотреть все три возможных варианта: результат значимо больше 25% (значит, вы добились успеха), значимо меньше 25% (что указывает на неудачу) и незначимо отличается от 25% (что указывает на недостаток информации для определенного вывода). Определите необходимую для этого случая двустороннюю гипотезу и выполните проверку на уровне 0,05.

в) Для двусторонней проверки кратко опишите результат, возможную ошибку и влияние результата на вашу вашей маркетинговую стратегию.

28. Вы проводите аудиторскую проверку, чтобы определить, являются ли ошибки в записях деловых счетов "ощутимыми ошибками". Для каждого счета есть заявленный в отчете остаток, аккуратность которого можно проверить только с помощью тщательного и дорогостоящего анализа; ошибка счета определяется как разность между заявленным в отчете остатком и действительным остатком счета. Отметим, что ошибка равна нулю тогда, когда отчет о счете составлен верно. Практически для данной ситуации (при наличии 12 000 счетов) *общая ошибка* будет ощутимой только тогда, когда она составит по крайней мере $5 000. Средняя ошибка для 250 случайно отобранных счетов составила $0,25 со стандартным отклонением $193,05. Поскольку речь может идти о вашей репутации аудитора, вы хотите иметь полную уверенность, прежде чем сделать заявление о том, что общая ошибка не является существенной.

а) Исходя из вашей выборки найдите оценку общей ошибки и сравните ее с размером существенной ошибки.

б) Сформулируйте нулевую и альтернативную гипотезы для односторонней проверки средней ошибки одного счета в генеральной совокупности и объясните, почему в данном случае можно использовать одностороннюю проверку.

в) Сформулируйте подходящее утверждение на основе одностороннего 95% доверительного интервала для средней ошибки одного счета в генеральной совокупности.

г) Вычислите t-статистику.

д) Какую гипотезу следует принять в результате односторонней проверки на уровне 5%?

е) Кратко опишите результаты этой аудиторской проверки.

29. Моющее средство для посудомоечных машин расфасовывают в упаковки весом 24 унции. Хотя вес пакетов различается, ваша цель обеспечить средний вес каждого выпущенного пакета на уровне немного больше, чем 24 унции. Случайная выборка 100 пакетов моющего средства из произведенной сегодня продукции дала следующие результаты: средний вес равен 24,23 унции со стандартным отклонением 0,15 унции.

а) Установите ρ-значение (либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$) для односторонней проверки гипотезы о том, что средний вес одной упаковки в генеральной совокупности превышает вес, указанный на упаковке.

б) Кратко опишите итоги проверки и ее результаты.

в) Вы сделали сильный или слабый вывод? Почему?

30. Чаще ли служащие берут больничные листы в последний год перед выходом на пенсию? У них может быть стимул поступать таким образом, если их общий оплаченный отпуск по болезни (разрешенное количество полностью оплаченных пропущенных по болезни рабочих дней) почти исчерпан. Действительно, такая ситуация наблюдается с государственными служащими. Этот вопрос был рассмотрен в статистических материалах Центрального финансово-контрольного управления США (U.S. General Accounting Office — GAO). В исследовании сделан следующий вывод: "Может ли быть так, что временное увеличение количества оплаченных по болезни дней — это просто отклонение выборки GAO, а не симптом желания уклониться от работы? Чтобы разобраться с этим вопросом, мы отмечаем, что 714 пенсионеров в выборке GAO в среднем были на больничном 30 дней в последний год перед выходом на пенсию вместо "ожидаемых" 14 дней. Иначе говоря, из 251 рабочего дня в году (среднее количество рабочих дней для федеральных служащих) 12% рабочего времени будущие пенсионеры находились на больничном, вместо предусмотренных 5,6%.

[19] Voos P. B., *цитируемое произведение*.

Могло ли так получиться случайно? Статистика утверждает, что вероятность таких отклонений в столь большой выборке очень низка. Чтобы быть точным: один из 200 000".

а) Определите генеральную совокупность и выборку.

б) Сформулируйте гипотезу, которую следует проверить, в терминах процента времени болезни.

в) Определите ρ-значение.

г) Какая гипотеза (если она есть) была отвергнута? Какая была принята?

д) Насколько статистически значим результат?

31. Взаимные инвестиционные фонды, которые вкладывают средства только в "социально значимые" фирмы, в первой половине 1988 г. сработали хорошо. Хотя среднее значение нормы прибыли для всех взаимных инвестиционных фондов составило 9,41%, норма прибыли для этих фондов была такой, как показано в табл. 10.7.3.

а) Можно ли рассматривать эти фонды действительно как случайную выборку из всех взаимных инвестиционных фондов, или наблюдаемая для них средняя норма прибыли — явление специфическое? Чтобы решить этот вопрос, выполните обычным образом проверку гипотезы на уровне 5%.

б) Установите ρ-значение этого теста (либо как $\rho > 0{,}05$, либо как $\rho < 0{,}05$, либо как $\rho < 0{,}01$, либо как $\rho < 0{,}001$). В частности, является ли результат высоко значимым?

в) Определите основную гипотезу и допущения, сделанные в п. "а".

г) При данных допущениях проверка гипотезы позволяет сделать ясное и корректное утверждение. Однако реалистичны ли данные допущения? Убедитесь, что фонды независимы (следует отметить, что некоторые из этих фондов являются частью одной и той же группы).

д) Можно утверждать, что если вы располагаете большим количеством обозревателей и аналитиков, отслеживающих работу каждой группы взаимных инвестиционных фондов, то, скорее всего, вы увидите крайние значения (примеры наиболее успешной и наименее успешной деятельности). Если имеется 100 групп, то можно ожидать, что приблизительно 5 из них будут значимо отличаться от общего среднего, если проверка производится на уровне 5%. Учитывая это, будет ли проверка (на основе данной выборки) реалистичной или следует относиться к полученному результату с осторожностью?

Таблица 10.7.3. Результаты социально значимых инвестиционных фондов

Ariel Growth	34,31%	Parnassus	36,45%
Calvert Bond	4,60	Pax World	7,15
Calvert Growth	8,50	Pioneer	15,63
Calvert Equity	13,20	Pioneer II	17,19
Dreyfus Third Century	14,87	Pioneer Bond	3,85
New Alternatives	18,89	Pioneer III	26,72

W. Stevens "Socially Aware Investing Turns Profitable". *The Wall Street Journal*, July 29, 1988, p. 27. Норма прибыли приведена по состоянию на 14 июля 1988 г.

32. В 1998 году мировые рынки инвестиций были очень нестабильны. В табл. 10.7.4 приведены значения нормы прибыли взаимных инвестиционных фондов закрытого типа, специализирующихся на получении дохода из международных источников.

а) Значимо ли отличается прибыль таких закрытых инвестиционных фондов, как группы, от средней прибыли 2,59% всех взаимных инвестиционных фондов во всем мире за тот же период времени? Если да, то значимо лучше или значимо хуже результаты этих закрытых взаимных инвестиционных фондов? При вычислениях можно принять, что средняя прибыль измерена без случайности.

б) Значимо ли отличается прибыль таких закрытых инвестиционных фондов, как группы, от средней прибыли −26,83% всех взаимных инвестиционных фондов на финансовых рынках стран с развивающейся экономикой за тот же период времени? Если это так, то значимо лучше или значимо хуже эти закрытые взаимные инвестиционные фонды? При вычислениях можно принять, что средняя прибыль измерена без случайности.

33. В прошлом году брокер, управляющий вашим портфелем ценных бумаг, добился прибыли 18,3%. В недавно опубликованной статье описан анализ работы выборки 25 брокеров, работающих в этой же сфере, который де-

Таблица 10.7.4. Результаты работы взаимных инвестиционных фондов закрытого типа, специализирующихся на получении доходов из международных источников: рыночная прибыль за год

Фонд	Норма прибыли, %	Фонд	Норма прибыли, %
AMC Mgd $-x	−27,7	Global Partners-x	−16,7
Alliance Wld $	−17,1	Kleinwort Aust	−5,9
Alliance Wld $ 2	−27,0	Morg St Em Debt-x	−24,9
BlckRk North Am-x	3,9	Morgan St Glbl-x	−27,3
Dreyfus Str Govt	4,0	Salomon SBG-x	−0,6
Emer Mkts Float	−19,7	Salomon SBW-x	−18,9
Emer Mkts Inc-x	−18,4	Scudder Glbl High Inc-x	−53,8
Emer Mkts Inc II-x	−16,9	Strategic Gl Inc	5,8
First Aust Prime-x	−5,3	Templeton Em Inc	−12,8
First Commonwlth-x	−3,5	Templtn Gl Govt	−1,1
Global HI Inc $	−10,7	Templtn Glbl Inc	2,2
Global Income Fund-x	−17,3	Worldwide $Vest-x	−48,2

Взято из "Quarterly Closed-End Funds Review", *The Wall Street Journal*, 1999, January 7, p. R14. Полные данные приведены в "Mutual-Fund Performance Yardsticks," p. R3.

монстрирует, что средняя прибыль в этой выборке составила 15,2% со стандартным отклонением 3,2% (измерено в процентных единицах).

а) Чтобы проверить, значимо ли превзошел ваш брокер эту группу, определите идеализированную совокупность и гипотезу, которую нужно проверить. В частности, вы будете проверять гипотезу относительно *среднего значения* или относительно *нового наблюдения*?

б) Определите стандартную ошибку предсказания.

в) Определите двусторонний 95% интервал предсказания для нового наблюдения.

г) Превзошел ли ваш брокер эту группу брокеров?

д) *Значимо* ли превзошел ваш брокер эту группу брокеров?

е) Определите t-значение и ρ-значение (либо $\rho > 0{,}05$, либо $\rho < 0{,}05$, либо $\rho < 0{,}01$, либо $\rho < 0{,}001$) для двусторонней проверки.

34. В прошлом году вы продали 3 834 новых автомобиля и получили в среднем 129,2 рекламации (необходимость исправить некоторые поломки в течение гарантийного срока) на один проданный новый автомобиль, со стандартным отклонением 42,1 рекламации. В этом году вы составили программу обеспечения качества для устранения некоторых из этих проблем еще до продажи автомобилей. Пока в этом году вы продали 74 автомобиля и получили в среднем только 93,4 рекламации на один проданный новый автомобиль со стандартным отклонением 37,7.

а) Какой метод проверки гипотезы необходимо использовать, чтобы посмотреть, работает ли ваша новая программа обеспечения качества?

б)Определите генеральные совокупности, выборки и гипотезы.

в) Выполните двустороннюю проверку на уровне 5% и опишите результат.

35. Почему фирмы меняют владельцев? Одной из возможных причин является то, что новые владельцы будут более эффективно управлять работой фирмы, чем старое руководство. Эта теория приводит к гипотезам, которые могут быть проверены. Например, эта теория предполагает, что при слиянии компаний производительность должна увеличиться и что фирмы, меняющие владельца, обычно имеют более низкую производительность, чем остальные фирмы. В течение ряда лет исследовалась производительность в тех фирмах, которые меняли владельцев, и в тех, которые не меняли. В частности, отчет об этом исследовании содержит следующее:

"Эти цифры показывают очень четкую картину. Фабрики, которые сменили владельцев... были менее эффективными... чем предприятия, на которых руководство не менялось... Но различия... (после смены владельца) сглаживались... Это означает, что производительность фирм... сменивших владельца, до этой смены была ниже производительности фирм... не менявших владельца, и постоянно падала, а затем (после смены владельца) начала, хотя и медленно, увеличиваться. За исключением одного случая все значения разностей производительности являются *статистически высоко значимыми*".

а) Объясните, что означает в цитированном отрывке выражение "статистически высоко значимые"?

б) Рассмотрим сравнение средней производительности (в момент смены владельца) тех фирм, которые сменили владельца, со средней производительностью тех фирм, которые не сменили владельца. Определите все, что необходимо для проверки гипотезы, и, в частности гипотезу, данные выборки, метод проверки и сделанные предварительные допущения.

в) Один из результатов был описан следующим образом: "в момент смены владельца уровень производительности был на 3,9% ниже по сравнению с фабриками, которые не сменили владельцев. Значение t-статистики равно 9,10". Исходя из данной информации выполните проверку гипотезы и дайте свое заключение.

г) Зачем те, кто проводили это исследование, использовали проверку статистических гипотез? Что они получили дополнительно по сравнению с простым наблюдением и описанием различий производительности в имеющихся данных?

36. Существует мнение, что стресс, возникающий, когда человек говорит неправду, может быть измерен. Для шести человек с помощью детектора лжи был измерен уровень стресса как тогда, когда они давали правдивые ответы на поставленные вопросы, так и тогда, когда они говорили неправду. Результаты этих измерений приведены в табл. 10.7.5.

а) Был ли уровень стресса у каждого из шести человек выше при лживом ответе, чем при честном ответе?

б) Определите средние уровни стресса для честных и лживых ответов. Определите среднее значение изменения уровня стресса (уровень стресса при лживом ответе минус уровень стресса при правдивом).

в) Определите соответствующую стандартную ошибку для среднего разности. В частности, в данной ситуации мы имеем дело с зависимыми или с независимыми выборками?

г) Определите 95% двусторонний доверительный интервал для среднего разности уровней стресса.

д) Проверьте, значимо ли отличаются средние значения уровня стресса.

Таблица 10.7.5. Уровень стресса во время ответов на вопросы

Человек	Честный ответ	Лживый ответ
1	12,8	13,1
2	8,5	9,6
3	3,4	4,8
4	5,0	4,6
5	10,1	11,0
6	11,2	12,1

Если они отличаются значимо, то уровень стресса при ложном ответе значимо выше или значимо ниже?

е) Кратко опишите результаты этой проверки. Особенно отметьте, является ли полученный результат заключением именно об этих шести людях или о некоторой другой группе? Как можно определить значимую разницу, если во время ложного ответа у одних людей стресс повышается, а у других наоборот понижается?

37. Группа экспертов провела оценку двух лучших сортов вин, выпускаемых вашим винным заводом. Оценка проводилась по 20-балльной шкале (чем выше балл, тем лучше вино). Результаты приведены в табл. 10.7.6.

а) Это зависимые или независимые выборки? Почему?

б) Определите среднюю оценку для каждого типа марочного вина и среднее разности оценок ("Шардене" минус "Каберне Совиньон").

в) Определите соответствующую стандартную ошибку для среднего разности оценок.

г) Определите 95% двусторонний доверительный интервал для среднего разности оценок.

д) Проверьте, значимо ли отличаются средние оценки. Если они отличаются значимо, то какая марка вина лучше?

е) Кратко изложите результаты этой проверки.

38. Для определения конкурентоспособности вашей продукции проведены испытания надежности вашего изделия и сравнение его с аналогичным изделием ваших ближайших конкурентов. Во время испытаний правила эксплуатации нарушались таким образом, что амортизация каждого изделия за день была приблизительно эквивалентна его амортизации за год нормальной работы. В табл. 10.7.7 приведены данные о том, как долго смог выдержать каждый экземпляр.

а) Определите среднее время работы до поломки для ваших изделий и для изделий конкурентов. Определите разность средних (ваши изделия минус изделия ваших конкурентов).

б) Определите соответствующую стандартную ошибку для разности средних. В частности, являются ли эти данные зависимыми или независимыми? Почему?

Таблица 10.7.6. Оценка качества вина (в баллах)

Эксперт	"Шардене"	"Каберне Совиньон"	Эксперт	"Шардене"	"Каберне Совиньон"
1	17,8	16,6	6	19,9	18,8
2	18,6	19,9	7	17,1	18,9
3	19,5	17,2	8	17,3	19,5
4	18,3	19,0	9	18,0	16,2
5	19,8	19,7	10	19,8	18,6

в) Определите 99% двусторонний доверительный интервал для разности средних значений надежности.

г) Проверьте на уровне 1%, есть ли значимое отличие надежности ваших изделий по сравнению с изделиями конкурентов.

д) Определите ρ-значение для разницы в надежности (либо как $\rho > 0{,}05$, либо $\rho < 0{,}05$, либо $\rho < 0{,}01$, либо $\rho < 0{,}001$).

е) Напишите краткое резюме, информацию из которого можно использовать в брошюре, рекламирующей ваши изделия.

39. Уход за маленьким ребенком является жизненно важной проблемой для работающих родителей. В табл. 10.7.8 приведены данные о месячной стоимости ухода за одним ребенком в выборке центров по дневному уходу за детьми в округе Северный Сиэтл.

40. Район Лаурелхерст считается наиболее привлекательным для проживания, поэтому цены на жилье в этом районе выше. Выполните одностороннюю проверку на уровне 5% гипотезы о том, что и стоимость ухода за ребенком в этом районе выше.

Рекламный опрос шести случайно отобранных людей в каждом из двух го+родов отражает уровень предпочтения нового изделия каждым из опрошенных (табл. 10.7.9).

а) Это задача для двух зависимых или двух независимых выборок?

б) Определите средний уровень предпочтения для каждого города.

в) Определите стандартную ошибку разности между средними уровней предпочтения. (Отметим, что речь идет о малых выборках.)

Таблица 10.7.7. Количестве дней работы изделия до поломки

Ваши изделия	Изделия ваших конкурентов
1,0	0,2
8,9	2,8
1,2	1,7
10,3	7,2
4,9	2,2
1,8	2,5
3,1	2,6
3,6	2,0
2,1	0,5
2,9	2,3
8,6	1,9
5,3	1,2
	6,6
	0,5
	1,2

г) Определите 95% двусторонний доверительный интервал для разности средних значений предпочтений в этих двух городах (Грин Бэй минус Милуоки).

д) Проверьте, является ли явно заметная разность в предпочтении статистически значимой на уровне 5%.

41. Одно и то же изделие изготавливают по новой и старой технологиям. В табл. 10.7.10 приведены данные по уровню брака изделий для каждой из технологий, измеренные для некоторого периода времени.

а) Оцените, насколько уменьшится уровень брака при переходе от старой технологии к новой?

б) Какой будет стандартная ошибка ответа на вопрос в п. "а"?

Таблица 10.7.8. Величина месячной стоимости ухода за одним ребенком в округе Северный Сиэтл (в долларах)

Район Лаурелхерст	За пределами района Лаурелхерст
400	500
625	425
440	300
550	350
600	550
500	475
	325
	350
	350

Таблица 10.7.9

Милуоки	Грин Бэй
3	4
2	5
1	4
1	3
3	2
2	4

Таблица 10.7.10

	Старая технология	Новая технология
Средний уровень брака	0,047	0,023
Стандартное отклонение	0,068	0,050
Размер выборки (дни)	50	44

в) Ваша фирма будет заинтересована в переходе на новую технологию только тогда, когда убедится, что такой переход улучшит качество. Сформулируйте нулевую и альтернативную гипотезы для этой ситуации.

г) Определите соответствующий односторонний 95% доверительный интервал для снижения уровня брака в генеральной совокупности в долгосрочной перспективе.

д) Является ли улучшение качества продукции (согласно оценке в п. "а") статистически значимым?

42. Чтобы решить, с кем из ваших двух нынешних поставщиков стоит увеличить объем контракта в следующем году, вы изучили случайные выборки пластиковых ящиков, полученных от каждого из них. В табл. 10.7.11 приведены данные нескольких измерений (большее число указывает на более высокое качество).

а) Определите среднее качество продукции для каждого из поставщиков.

б) Определите стандартное отклонение качества для каждого из поставщиков.

в) Определите разность средних оценок качества (International минус Custom) и его стандартную ошибку.

г) Определите двусторонний 95% доверительный интервал для разности оценок качества.

д) Есть ли значимое различие в качестве продукции этих двух поставщиков? Как вы это определили?

43. Рассмотрите две выборки значений веса конфет до и после вмешательства в процесс производства из задачи 11 главы 5.

а) Мы имеем дело с зависимыми или независимыми данными?

б) Найдите 95% доверительный интервал для разности средних значений веса одной конфеты в генеральной совокупности (вес после вмешательства минус вес до вмешательства).

в) Оказывает ли вмешательство в процесс производства значимое влияние на вес конфеты? Как вы это определили?

44. Из 983 деталей, произведенных отделением вашей фирмы в Детройте на прошлой неделе, было забраковано 135. За тот же самый период из 1085

Таблица 10.7.11

Качество продукции поставщиков	
Custom Cases Corp.	International Plastics, Inc.
54,3	93,6
58,8	69,7
77,8	87,7
81,1	96,0
54,2	82,2
78,3	

деталей, изготовленных вашим отделением в Канзас-Сити, было забраковано 104 детали.

а) Определите процент брака в каждом из отделений и сравните эти значения.

б) Определите разность между этими двумя процентами (Детройт минус Канзас-Сити) и дайте интерпретацию полученного значения.

в) Определите стандартную ошибку разности, используя формулу для большой выборки.

г) Определите 95% доверительный интервал для этой разности.

д) Исходя из уровня брака определите, значимо ли отличается качество продукции этих двух филиалов.

45. Вы анализируете результаты опроса потребителей изделия, измеренные по 10-балльной шкале. Для 130 потребителей, которые описали себя как "коммуникабельные", средняя оценка составила 8,36 со стандартным отклонением 1,82. Для 218 "застенчивых" потребителей средняя оценка составила 8,78 со стандартным отклонением 0,91.

а) Проверьте, значимо ли различие между оценками "коммуникабельных" и "застенчивых" потребителей.

б) Изложите результаты проверки в терминах ρ-значения (либо $\rho > 0{,}05$, либо $\rho < 0{,}05$, либо $\rho < 0{,}01$, либо $\rho < 0{,}001$).

46. Решите предыдущую задачу для другого вида изделий. Для 142 "коммуникабельных" потребителей средняя оценка составила 7,28 со стандартным отклонением 2,18. Для 277 "застенчивых" потребителей средняя оценка составила 8,78 со стандартным отклонением 1,32.

47. Решите задачу 45 для еще одного вида изделий. Для 158 "коммуникабельных" потребителей средняя оценка составила 7,93 со стандартным отклонением 2,03. Для 224 "застенчивых" потребителей средняя оценка составила 8,11 со средним отклонением 1,55.

48. Вы определили, что в чашке кофе содержится только 72,8 мг кофеина. Проверьте (на уровне 5%), могут ли использованные в данном случае кофейные зерна принадлежать той же генеральной совокупности, что и зерна, данные о которых использованы в задаче 41 главы 9.

Упражнения с использованием базы данных

Рассмотрим базу данных служащих из приложения А. Считайте этот набор данных случайной выборкой из более крупной генеральной совокупности служащих.

1. Значимо ли средняя заработная плата за год отличается от \$40 000? 
2. Вы хотели бы утверждать, что в генеральной совокупности служащих средний стаж работы значимо превышает пять лет. Можете ли вы доказать это утверждение?
3. Проверьте, значимо ли различается доля в 50% у мужчин и женщин.
4. Проверьте, различаются ли в генеральной совокупности средняя заработная плата за год мужчин и женщин.

5. Проверьте, различаются ли в генеральной совокупности средний возраст мужчин и женщин.
6. Проверьте, значимо ли отличается средний размер заработной платы служащих квалификации A от среднего размера заработной платы служащих квалификаций B и C, вместе взятых.
7. Проверьте, отличается ли в генеральной совокупности средний возраст служащих квалификации A от среднего возраста служащих квалификаций B и C, вместе взятых.

Проекты

1. Выберите какой-нибудь процесс принятия решения, связанный с вашей работой или бизнесом, который можно осуществить на основе анализа данных.

а) Опишите нулевую и альтернативную гипотезы.

б) Вычислите (или используйте в целях обучения некоторое предположение) соответствующую оценку и ее стандартную ошибку.

в) Постройте доверительный интервал.

г) Проверьте гипотезу.

д) Поясните и опишите полученные результаты.

2. Найдите любое сообщение (в Internet, в газете, журнале, сообщениях радио или телевидения), которое содержит заключение на основе анализа данных.



а) Определите нулевую и альтернативную гипотезы.

б) Определите, насколько это возможно из имеющейся информации, генеральную совокупность и выборку. Как вы считаете, не пропущена ли какая-либо важная информация?

в) Каков результат проверки гипотезы в сообщении?

г) Является ли сделанное в сообщении заключение слабым или сильным?

д) Обсудите и интерпретируйте утверждения, сделанные в сообщении.

Ситуация для анализа

Так много рекламы, так мало времени

Подошло время принятия решения, ставки очень высоки. Ввиду астрономических цен за минуту показа рекламы на телевидении стоит провести некоторую предварительную работу, чтобы не тратить напрасно деньги. В частности, вы помогали руководству в подготовке 22 рекламных роликов о товарах личной гигиены, хотя только несколько из них будет показано широкой публике. Все рекламные ролики необходимо было проверить и оценить, используя ответы представителей возможных покупателей, каждый из которых был отобран случайным образом, каждому был показан один рекламный ролик и заданы вопросы до и после просмотра. Обобщенная оценка по шкале от 1 до 10 баллов, включающая как запоминаемость рекламы, так и ее убедительность, была получена для каждого из покупателей, попавших в выборку.

На вашей фирме рекламу, как правило, оценивают, используя среднее значение обобщенной оценки, и набравший наибольшее количество баллов рекламный ролик запускают на общенациональный канал телевидения. Однако недавно принято решение использовать проверку статистической гипотезы, чтобы убедиться в том, что показываемый рекламный ролик (или ролики) оценивается значимо выше, чем минимальная оценка 3,5 балла.

На этот раз все выглядит достаточно очевидно для двух наилучших рекламных роликов, получивших оценку значимо выше минимума. Решение принимается просто: фаворитом является ролик "Пикник на лоне природы" для показа по телевидению в наилучшее время и ролик "Перерыв на кофе" в качестве альтернативы ему. Ниже приведены результаты оценивания, упорядоченные по убыванию среднего значения обобщенной оценки. Количество потребителей, которым показали рекламу, обозначено как n. p-значение для односторонней проверки гипотезы о сравнении с заданным значением 3,5 вычислялось отдельно для каждого рекламного ролика.

Рекламный ролик	n	Среднее	Стандартное отклонение	Стандартная ошибка	t	p
Пикник на природе	49	3,95	0,789	0,113	3,985	0,0001
Перерыв на кофе	51	3,70	0,744	0,104	1,921	0,0302
Юбилей	51	3,66	0,934	0,131	1,214	0,1153
Океанский бриз	49	3,63	0,729	0,104	1,255	0,1078
Друзья за игрой	56	3,62	0,896	0,120	0,969	0,1683
Теннисный матч	56	3,60	0,734	0,098	1,037	0,1521
Совместная прогулка	51	3,57	0,774	0,108	0,687	0,2476
Плавательный бассейн	52	3,56	0,833	0,116	0,532	0,2984
Посещение магазина	49	3,54	0,884	0,126	0,355	0,3619
Бег трусцой	47	3,54	0,690	0,101	0,423	0,3372
Семейная сценка	54	3,54	0,740	0,101	0,404	0,3438
Уединенный домик в горах	49	3,53	0,815	0,116	0,298	0,3836
Прохладный и удобный	52	3,52	0,780	0,108	0,195	0,4229
Вместе за чашкой кофе	53	3,52	0,836	0,115	0,148	0,4415
Городской пейзаж	47	3,51	0,756	0,110	0,058	0,4770
Друзья за работой	53	3,50	0,674	0,093	0,020	0,4919
Плаванье под парусом	48	3,49	0,783	0,113	–0,055	0,5219
Оазис в пустыне	55	3,48	0,716	0,097	–0,226	0,5890
Празднование дня рождения	50	3,48	0,886	0,125	–0,175	0,5693
Завтрак на отдыхе	53	3,45	0,817	0,112	–0,437	0,6681
Домой с работы	55	3,35	0,792	0,107	–1,430	0,9207
Ветреный день	47	3,34	0,678	0,099	–1,593	0,9410

После некоторых размышлений у вас появляются следующие мысли. Поскольку вы действительно хотите понять, на чем основано принимаемое решение, и помните из давно прослушанного вами курса статистики об ошибках при проверке гипотез, вы удивлены. Вероятность совершить ошибку I рода составляет 0,05, поэтому вы ожидаете, что относительно 1 из 20 рекламных роликов будет принято решение, что он значимо хорош, хотя это и не так. Таким образом, может оказаться так, что не будет найдено ни одного значимо хорошего, в то время как реально может быть даже больше одного значимо хорошего ролика.

Вы продолжаете размышлять: может ли так случиться, что решение будет принято на основе чистой случайности? Может ли случиться так, что потребители оценят эти ролики в среднем одинаково хорошо? Может ли случиться так, что все, что у нас есть, — это всего лишь случайный выбор определенной группы потребителей для оценки рекламы?

Вы решаете запустить компьютерную имитационную модель, установив для всех рекламных роликов в качестве среднего генеральной совокупности в точности 3,5 балла. Нажав 10 раз кнопку повторного вычисления, вы видите в электронной таблице, что три раза не было ни одного значимого рекламного ролика, 5 раз значимым оказался один ролик, один раз было два и один раз — три значимых ролика. Обычно каждый раз значимым оказывался другой ролик. Более того, случайно смоделированные результаты выглядят гораздо реальнее тех, которые вы собираетесь использовать для принятия решения.

Вопросы для обсуждения

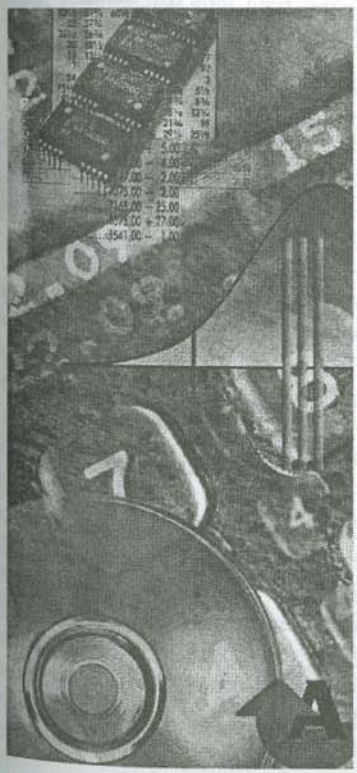
1. Выберите два рекламных ролика, один значимый, другой нет. Используя значения среднего, стандартной ошибки и размер выборки, проверьте значимость. Можно ли в этом случае использовать одностороннюю проверку?
2. Если предполагалось, что ошибка I рода контролируется на уровне 5%, то как могло произойти, что при компьютерном моделировании ошибка I рода наблюдались в 70% случаев?
3. Может ли быть так, что в исследовании, в котором 2 из 22 роликов оказались значимыми, ни один из рекламных роликов не заслуживает внимания?
4. Какова ваша интерпретация эффективности рекламы в этом исследовании. Что бы вы порекомендовали в этой ситуации?

Регрессия и временные ряды

В этой части...

Глава 11. "Корреляция и регрессия: измерение и прогнозирование взаимосвязей"
Глава 12. "Множественная регрессия: прогнозирование одного фактора на основе нескольких других"
Глава 13. "Составление отчетов: представление результатов множественной регрессии"
Глава 14. "Временные ряды: анализ изменений во времени"

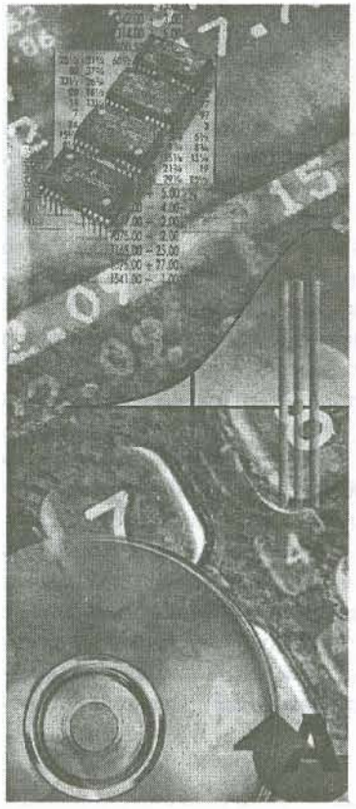
Итак, вы уже познакомились с основами статистики: вы знаете, как анализировать данные, вычислять и интерпретировать вероятности, как получить случайную выборку и сделать статистический вывод. Теперь наша задача заключается в том, чтобы применить все эти концепции для выявления различных взаимосвязей, скрывающихся в сложных ситуациях реальной жизни. В главе 11 будет показано, как статистика позволяет выявить взаимосвязь между двумя факторами на основе двумерной совокупности данных в виде двух столбцов чисел. Показатель *корреляции* покажет, насколько сильна эта взаимосвязь, а *регрессия* позволит прогнозировать один фактор на основе другого. Самым важным статистическим методом, возможно, является *множественная регрессия*, речь о которой пойдет в главе 12. Именно множественная регрессия позволяет использовать все имеющиеся у вас факторы для предсказания (т.е. снижения уровня неопределенности) некоторого важного, но неизвестного значения. Поскольку *общение* представляет собой весьма важный инструмент бизнеса, в главе 13 речь пойдет о том, как с максимальной эффективностью довести до сведения других людей полезную информацию, полученную вами в результате анализа методом множественной регрессии. Несмотря на неизменность базовых концепций, для *анализа временных рядов*, представленного в главе 14, требуются такие новые способы применения статистических методов, которые позволяют извлечь дополнительную информацию, содержащуюся во временной последовательности наблюдений.



ГЛАВА 11

Корреляция и регрессия: измерение и прогнозирование взаимосвязей

Окружающий нас мир полон всевозможных взаимосвязей: между отношением к труду и производительностью, между корпоративной стратегией и долей рынка, между вмешательством государства и состоянием экономики, между объемом выпускаемой продукции и затратами, между сбытом и доходами и т.п.



До сих пор нас интересовали главным образом такие статистические характеристики, как среднее значение и отклонение, которых обычно бывает достаточно, когда приходится иметь дело с *одномерными* данными (т.е. лишь с *одним* измерением — например, заработной платой) о каждой элементарной единице (например, о служащем). Когда вы имеете дело с *двумерными* данными (например, заработной платой и образованием), всегда есть возможность изучать каждое измерение по отдельности — как часть одномерной совокупности данных. Однако реальную отдачу можно получить лишь от совместного изучения обоих измерений, что дает возможность выявить взаимосвязь между ними.

Изучая взаимосвязи в двумерных данных, следует всегда помнить о следующих трех основных целях.

Первая. *Описание и понимание взаимосвязи.* Это самая общая цель, обеспечивающая получение базовой информации, с помощью которой можно лучше понять истинное устройство окружающего нас мира. При изучении сложной системы очень важно знать, какие факторы наиболее тесно взаимодействуют друг с другом, а какие вообще не оказывают влияния друг на друга. Знание этой информации может оказать значительную помощь в долгосрочном планировании и принятии других стратегических решений.

Вторая. *Прогнозирование и предсказание нового наблюдения.* Понимание некоторой взаимосвязи может позволить использовать информацию об одном из измерений для более качественного предсказания другого измерения. Если, например, вам известно, что в этом квартале количество заказов на продукцию увеличилось, можно ожидать и увеличения объема сбыта. Если вы проанализировали взаимосвязь между количеством заказов и объемами сбыта в прошлом, у вас есть все шансы сделать достоверный прогноз сбыта на будущее, основываясь на текущем количестве заказов.

Третья. *Регулирование и управление процессом.* Когда вы *вмешиваетесь* в какой-либо процесс (например, регулируете уровень производства, вводя некоторые технологические изменения или новый тип обслуживания), необходимо определить объем этого вмешательства. Если существует непосредственная взаимосвязь между вмешательством и результатом и вы эту взаимосвязь понимаете, то такое знание может помочь вам выполнить оптимальное регулирование.

Двумерные данные могут иметь различную структуру; с некоторыми структурами работать легко, с другими — труднее. Исследование данных с помощью *диаграммы рассеяния* позволяет увидеть то, что находится за привычными статистическими характеристиками. Существуют два базовых инструмента, с помощью которых анализируют двумерные данные: *корреляционный анализ*, позволяющий оценить степень взаимосвязи между двумя факторами (если такая взаимосвязь вообще существует), и *регрессионный анализ*, показывающий, как можно предсказать или управлять одной из двух переменных с помощью другой. Проверка статистических гипотез позволяет оценить взаимосвязь, которая, как вам кажется, существует в изучаемых данных, и выяснить, является ли она значимой или может быть объяснена исключительно случайностью.

11.1. Исследование взаимосвязей с помощью диаграмм рассеяния и корреляций

Когда приходится иметь дело с двумерными данными, следует нарисовать *диаграмму рассеяния*, которая позволяет *увидеть* структуру. Так же как гистограмма отображает структуру одномерных данных (нормальное распределение, асимметрия, выбросы и т.д.), диаграмма рассеяния показывает вам все, что происходит с двумерными данными. Если ваши данные содержат какие-то проблемы (например, выбросы или какие-то неожиданные особенности), зачастую единственный способ их обнаружения состоит как раз в анализе соответствующей диаграммы рассеяния.

Корреляция является мерой силы взаимосвязи. Подобно всем статистическим характеристикам, корреляция одновременно и полезна, и ограничена. Если диаграмма рассеяния показывает, например, ярко выраженную *линейную* взаимосвязь (о которой мы вскоре поговорим подробнее) или отсутствие какой-либо взаимосвязи, то корреляция превосходно это отражает. Но если данные содержат определенные проблемы (о которых мы также поговорим ниже подробнее), такие как *нелинейная* взаимосвязь, *неодинаковая изменчивость*, наличие *групп* или *выбросов*, корреляция может вводить в заблуждение.

Сама по себе корреляция носит ограниченный характер, поскольку ее интерпретация зависит от типа взаимосвязи в данных. Вот почему столь большое значение придается диаграмме рассеяния: она либо подтверждает обычную интерпретацию корреляции, либо показывает наличие в данных определенных проблем, которые приводят к тому, что корреляция лишь вводит нас в заблуждение.

Диаграмма рассеяния демонстрирует взаимосвязь

Диаграмма рассеяния представляет каждое наблюдение (или элементарную единицу) в пространстве двух измерений, соответствующих двум факторам. Если одна переменная рассматривается как "причина", влияющая на другую переменную, она обозначается буквой *X* и ей соответствует горизонтальная ось. Реагирующая на это влияние переменная обозначается буквой Y, и ей соответствует вертикальная ось. Если нельзя четко указать, какая переменная оказывает влияние, а какая подвержена влиянию, то можно просто обозначить один фактор X, а другой — Y.

Диаграмма рассеяния для небольшой двумерной совокупности данных, представленной в табл. 11.1.1, показана на рис. 11.1.1. Поскольку принято считать, что затраченные усилия влияют на результаты, было бы вполне естественным отобразить число контактов с клиентами (т.е. затраченные усилия) на горизонтальной оси, а объем продаж (результат) — на вертикальной. Иногда бывает удобно пометить точки, как это сделано на рис. 11.1.1 (хотя иногда подобная разметка лишь загромождает диаграмму и мешает восприятию общей картины). Более привычный вид диаграммы рассеяния для этой совокупности данных показан на рис. 11.1.2.

На обоих рисунках представлена информация как о каждой отдельной переменной, так и о взаимосвязи между ними. Во-первых, распределение количества контактов (см. горизонтальную ось) находится приблизительно в диапазоне от 150 до 220, причем типичное значение равно приблизительно 170. Во-вторых, распределение объемов сбыта (см. вертикальную ось) находится в диапазоне приблизительно от $130 000 до $180 000, причем типичное значение равно приблизительно $150 000. Наконец, взаимосвязь между количеством контактов и объемом продаж оказалась положительной: точки на диаграмме выстраиваются снизу вверх при движении слева направо. Это свидетельствует о том, что сотрудники, имеющие больше контактов с клиентами (соответствующие точки расположены на диаграмме правее), обеспечили фирме и большие объемы сбыта (соответствующие точки расположены на диаграмме выше). Для данных в целом характерен такой рост, но это справедливо не для всех наблюдений. Это типично

Таблица 11.1.1. Результаты работы по итогам первого квартала

	Контакты	Объем продаж, дол.
Билл	147	126 300
Марта	223	182 518
Коллин	163	141 775
Гэри	172	138 282

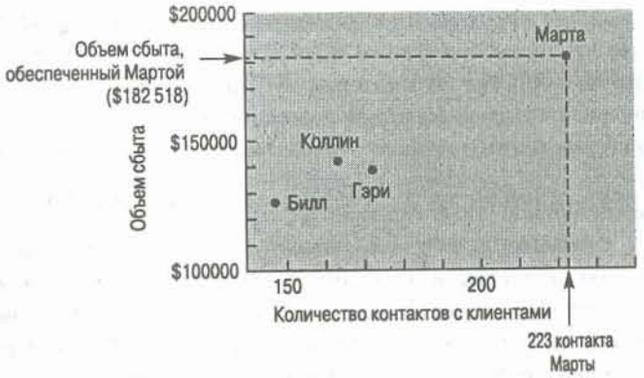


Рис. 11.1.1. Эта диаграмма рассеяния содержит по одной точке для каждой строки вашей двумерной совокупности данных. Каждая точка диаграммы имеет метку, свидетельствующую о ее "происхождении". Выделены выдающиеся достижения Марты — 223 контакта, результатом которых стал квартальный объем продаж на $182 518

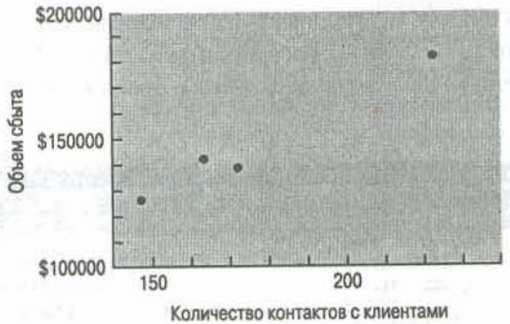


Рис. 11.1.2. Диаграмма рассеяния (как на предыдущем рисунке), но без дополнительной информации. Рисунок позволяет увидеть распределение количества контактов (вдоль горизонтальной оси), распределение объема продаж (вдоль вертикальной оси) и общее отношение возрастания объема продаж при росте количества контактов (т.е. точки на диаграмме поднимаются при движении вправо)

для статистического анализа, когда исследователя интересует тенденция, "общая картина"; выявленные при этом закономерности полезны, хотя данные и не соответствуют им идеально.

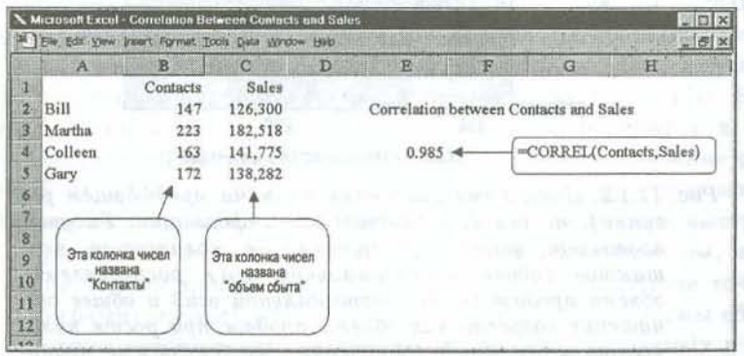
Корреляция характеризует силу взаимосвязи

Корреляция, или коэффициент корреляции, обозначаемый *r*, является числом в диапазоне от −1 до 1, характеризующим силу взаимосвязи в данных. Корреляция, равная 1, указывает на идеальную взаимосвязь в виде прямой линии,

причем более высокие значения одной переменной соответствуют идеально предсказуемым более высоким значениям другой переменной. Корреляция, равная –1, указывает на идеальную отрицательную взаимосвязь в виде прямой линии, причем одна переменная *уменьшается* с ростом другой.

Обычная интерпретация промежуточных корреляций в диапазоне от –1 до 1 заключается в том, что величина (абсолютное значение) корреляции указывает на "силу" взаимосвязи, а знак (положительный или отрицательный) указывает направление (увеличение или уменьшение). Обычная интерпретация корреляции, равной 0, заключается в том, что взаимосвязи нет, есть только случайность. Однако к такой интерпретации следует относиться с осторожностью, поскольку нелинейность и выбросы могут искажать обычную интерпретацию корреляции. Даже беглого взгляда на диаграмму рассеяния бывает достаточно, чтобы подтвердить или исключить подобные неприятные возможности. В табл. 11.1.2 показано, как интерпретировать корреляцию в каждом конкретном случае. Напомним, что корреляция показывает, насколько близко к указанной прямой линии располагаются точки на диаграмме. Она вовсе *не свидетельствует* о крутизне наклона этой линии.

Чтобы вычислить корреляцию средствами Excel®, можно воспользоваться функцией =CORREL() (=КОРРЕЛ()), указав названия двух столбцов чисел (можно использовать, например, команду меню Insert⇨Name⇨Define (Вставка⇨Имя⇨Присвоить), как показано на приведенном ниже рисунке. Функция CORREL вычисляет корреляцию между количеством контактов и объемом продаж, которая в данном случае равна 0,985.



Формула для вычисления коэффициента корреляции

Корреляция вычисляется на основе соответствующих данных с помощью достаточно простой формулы. Для проведения вычислений по этой формуле требуется довольно много времени, что, впрочем, не составляет большой проблемы при наличии компьютера или даже хорошего карманного калькулятора. Формула приведена скорее не для того, чтобы вы ее использовали, а для того, чтобы показать изнутри, как это все работает.

Таблица 11.1.2. Интерпретация коэффициента корреляции

Корреляция	Общепринятая интерпретация	Некоторые другие возможности
1	Идеальная положительная взаимосвязь. Все точки данных должны располагаться строго на прямой линии, направленной вверх и вправо	Отсутствуют
Близко к 1	Сильная положительная взаимосвязь. Точки данных довольно плотно сгруппированы (с небольшим случайным разбросом) вокруг прямой линии, направленной вверх и вправо	Точки данных располагаются строго на кривой, направленной вверх (нелинейная структура). Между точками данных взаимосвязи в основном нет, но один выброс данных (резко отклоняющаяся точка) исказил корреляцию. Корреляция искажена наличием в данных групп по-разному взаимосвязанных между собой объектов
Близко к 0, но положительно	Незначительная положительная взаимосвязь. Точки данных образуют случайное облако с незначительной ориентацией вверх и вправо	
0	Отсутствие взаимосвязи, совершенно случайное облако, не имеющее ориентации ни вверх, ни вниз при движении вправо	Точки данных располагаются строго на кривой, имеющей наклон вверх с одной стороны и наклон вниз с другой Точки данных располагаются строго на прямой линии, но один выброс данных (резко отклоняющаяся точка) исказил корреляцию. Корреляция искажена наличием в данных групп по-разному взаимосвязанных объектов
Близко к 0, но отрицательно	Незначительная отрицательная взаимосвязь. Точки данных образуют случайное облако с незначительной ориентацией вниз и вправо	Точки данных располагаются строго на кривой, направленной вниз (нелинейная структура). Точки данных в целом не образуют какой-либо структуры, но один выброс данных (резко отклоняющаяся точка) исказил корреляцию. Корреляция искажена наличием в данных групп по-разному взаимосвязанных между собой объектов
Близко к –1	Сильная отрицательная взаимосвязь. Точки данных плотно сгруппированы (с небольшим случайным разбросом) вокруг прямой линии, направленной вниз и вправо	
–1	Идеальная отрицательная взаимосвязь. Все точки данных должны располагаться строго на прямой линии, направленной вниз и вправо	Отсутствуют
Не определено	Точки данных располагаются строго на горизонтальной или на вертикальной линии	Недостаточно данных (менее n = 2 различных пар значений X и Y)

Формула для коэффициента корреляции включает двумерные данные, начиная с двух измерений ($X1$, $Y1$), сделанных для первого объекта, и заканчивая измерениями (Xn, Yn), сделанными для последнего объекта. Например, $X1$ может быть объемом сбыта компании IBM, а $Y1$ — чистым доходом IBM; Xn может быть объемом сбыта компании GM, а Yn — чистым доходом GM. Рассматривая каждый столбец чисел по отдельности, можно, например, вычислить обычное стандартное отклонение выборки только для значений X, т.е. найти S_x; аналогично, Sy представляет собой стандартное отклонение только лишь для

значений Y.[1] Формула для вычисления коэффициента корреляции также включает сумму попарных произведений значений X и Y, которая фиксирует их взаимозависимость, разделенную на $n - 1$ (как обычно поступают при вычислении стандартного отклонения).

Формула коэффициента корреляции

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}.$$

Слагаемые в числителе выражают взаимодействие двух переменных и определяют знак (положительной или отрицательной) корреляции. Если, например, между переменными существует сильная положительная взаимосвязь (увеличение одной переменной при увеличении второй), каждое слагаемое будет положительным числом: когда точка характеризуется высокими значениями X и Y, произведение будет положительным; когда точка характеризуется низкими значениями X и Y, произведение все равно будет положительным, поскольку оба сомножителя будут отрицательными (значения X и Y меньше соответствующих средних), а произведение двух отрицательных чисел является положительным числом. Аналогично, если между переменными существует сильная отрицательная взаимосвязь, все слагаемые в числителе будут отрицательными числами, что в результате дает отрицательное значение корреляции.

Знаменатель выражения для коэффициента корреляции просто нормирует числитель таким образом, что коэффициент корреляции оказывается легко интерпретируемым чистым (т.е. не имеющим размерности) числом в диапазоне от −1 до 1. Поскольку в числителе выражения присутствует произведение двух переменных, имеет смысл преобразовать его в "чистое" число, разделив на произведение сомножителей, включающих эти переменные. Если бы мы не выполнили такое деление, сам по себе числитель было бы трудно интерпретировать по причине возможной странности его единиц измерения. Если бы, например, X и Y измерялись в долларах, тогда числитель имел бы такую необычную размерность, как "доллары в квадрате".

Числитель выражения для коэффициента корреляции, который трудно интерпретировать из-за необычных единиц измерения, называется **ковариацией** X и Y. Несмотря на то что иногда он используется как самостоятельная характеристика (например, в теории финансов для описания совместного изменения курсов акций на двух биржах), удобнее пользоваться коэффициентом корреляции. Корреляция и ковариация представляют, по сути, одну и ту же информацию (при условии, что также известны отдельные стандартные отклонения), однако корреляция представляет эту информацию в более удобной форме.

Обратите также внимание на возможность поменять местами X и Y в этой формуле; иными словами, формула *симметрична* относительно X и Y. Таким образом, корреляция X с Y — это то же самое, что и корреляция Y с X, т.е. какая из двух переменных будет указана первой, значения не имеет. Это утверждение справедливо для корреляции, но не для регрессии (речь о которой пойдет в разделе 11.2).

[1] Обратите внимание, что S_x и S_y представляют собой стандартные отклонения, отражающие изменчивость *отдельных объектов*; их не следует путать со стандартными ошибками $S_{\bar{x}}$ и $S_{\bar{y}}$, отражающими изменчивость средних значений выборки — $\bar{X}$ и $\bar{Y}$ соответственно.

Различные типы взаимосвязей

В последующих разделах будут рассмотрены различные типы взаимосвязей, которые можно выявить, анализируя диаграмму рассеяния для двумерной совокупности данных. Для каждого типа взаимосвязей мы приведем по крайней мере по одному примеру, рассмотрим соответствующую диаграмму рассеяния, коэффициент корреляции и дадим некоторые комментарии.

Линейная взаимосвязь

Одни виды двумерных совокупностей данных легче анализировать, чем другие. Легче всего анализировать двумерные совокупности данных с *линейной взаимосвязью*. Эта взаимосвязь играет такую же особую роль для двумерных данных, как и нормальное распределение для одномерных данных. **Линейная взаимосвязь** проявляется в двумерной совокупности данных, если точки на диаграмме рассеяния случайным образом концентрируются вокруг прямой линии.[2] Эти точки могут концентрироваться довольно тесно, почти точно попадая на прямую линию, или быть разбросаны достаточно широко, образуя некоторое облако. Но такая взаимосвязь не должна быть криволинейной или воронкообразной, в данных не должно быть сильных выбросов (резко отклоняющихся значений).

Пример. *Рейтинги телевизионных программ компании Nielsen и "пиплметры"*

Вы, наверное, слышали о рейтингах телевизионных шоу, которые измеряет компания Nielsen. Поскольку расценки на телевизионную рекламу непосредственно зависят от величины телевизионной аудитории, эти рейтинги имеют особое значение для сетей телевидения и рекламодателей. Действительно, поскольку демонстрация одного 30-секундного рекламного ролика может стоить сотни тысяч долларов (плюс стоимость производства), даже небольшое изменение в объеме зрительской аудитории может иметь большое влияние на бюджет.

В течение приблизительно 30 лет рейтинги компании Nielsen основывались на записях в специальных дневниках, которые вели несколько тысяч американских семей. В этих дневниках семьи ежедневно указывали, какие передачи и в течение какого времени они смотрели. С появлением относительно недорогого электронного оборудования и компьютеров стало возможным использование так называемых "пиплметров", которые автоматически фиксируют соответствующую информацию. Переход от старой системы к новой оказался достаточно болезненным, поскольку два этих метода давали сильно различающиеся значения. Рассмотрим следующую ситуацию.

> Проблема заключается в том, что начальные испытания "пиплметров" свидетельствуют о том, что некоторые программы просматривает меньшее число людей, чем это следует из журналов, заполняемых вручную. Таким образом, рекламные агентства, которые ежегодно тратят на размещение рекламы на телевидении примерно 8 миллиардов долларов, хотели бы, чтобы рейтинги рекламы основывались исключительно на рейтингах "пиплметров", что, по их мнению, должно заставить телевизионные сети снизить цены... Важность рейтингов

[2] Говорят, что двумерная совокупность данных имеет *двумерное нормальное распределение*, если есть линейная взаимосвязь между переменными и, кроме того, если каждая из переменных имеет нормальное распределение. Более технически строгое определение требует также, чтобы для каждого значения X соответствующие ему значения Y были распределены нормально с одним и тем же стандартным отклонением.

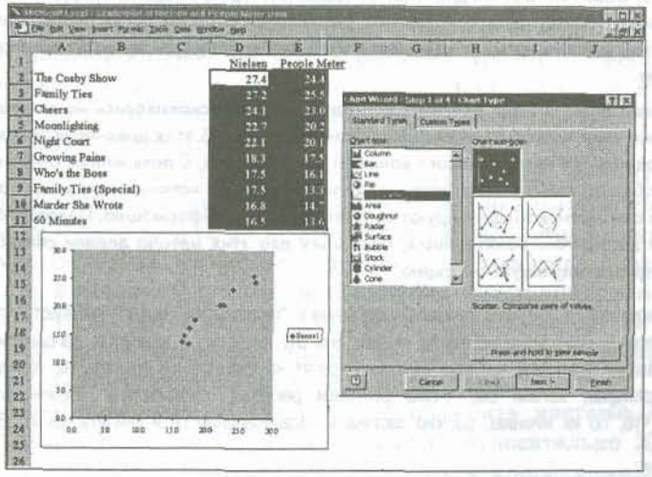
особенно очевидна для таких телешоу, как "Cosby" (стоимость показа 30-секундного рекламного ролика во время этого телешоу равняется примерно $380 000). По оценке с помощью "пиплметров" компании Nielsen в течение двух недель перед 23 ноября в число зрителей "Cosby" входило на 20% меньше женщин в возрасте от 18 до 49 лет и на 7% меньше мужчин из той же возрастной группе, чем следует из дневниковых записей за тот же период времени... Однако для других программ наблюдается обратная картина: показатели "пиплметров" превышают данные из дневниковых записей... Подобные расхождения могут оказывать значительное влияние на ход переговоров между рекламодателями и телевизионными сетями по поводу цен на рекламу... Nielsen стремится как можно скорее перейти к использованию пиплметров, причем эта поспешность в значительной степени объясняется конкуренцией со стороны AGB Television Research, Inc.[3]

В настоящее время Nielsen использует "пиплметры" в сочетании с записями в дневниках и по-прежнему остается авторитетной компанией в сфере исследований телевизионного рынка; однако Nielsen испытывает всевозрастающую конкуренцию со стороны Statistical Research, Inc.[4]

В *табл. 11.1.3 приведена двумерная совокупность данных, состоящая из индекса Nielsen (одна переменная) и индекса "пиплметра" (другая переменная), измеренных для каждого из n = 10 телешоу (элементарные единицы).*

Диаграмма рассеяния, представленная на рис. 11.1.3, соответствует линейной структуре, поскольку точки размещаются случайным образом вдоль прямой линии. Выявленная взаимосвязь является положительной, поскольку у телешоу с более высокими значениями индекса Nielsen, как правило, более высокие рейтинги "пиплметров". Высокая корреляция, $r = 0{,}974$, подтверждает факт существования сильной положительной (но не идеальной) связи. Здесь все же имеется некий элемент случайности, который может иметь немаловажное значение как для телесетей, так и для рекламодателей.

Чтобы воспользоваться Excel® для построения диаграммы рассеяния, нужно выбрать оба столбца чисел (разместив данные, соответствующие горизонтальной оси X, слева) и затем выбрать в меню команду Insert⇨Chart (Вставка⇨Диаграмма). Далее, в перечне типов диаграмм следует выбрать XY (Scatter) (Точечная). Продолжая выполнять последовательность шагов в Excel, можно создать диаграмму рассеяния в виде объекта рабочего листа. Ниже показано, как должно выглядеть начальное диалоговое окно, *после того как вы выберете соответствующие данные и начнете вставлять диаграмму, а также показан* окончательный вариант диаграммы на рабочем листе.



[3] Barnes P. and Lipman J. "Networks and Ad Agencies Battle over Estimates of TV Viewership", *The Wall Street Journal*, 1987, January 1, p. 25.

[4] Pope K. "Networks to Launch a Rival to Nielsen Service", *The Wall Street Journal*, 1998, August 3, p. B1.

Таблица 11.1.3. Телевизионные рейтинги

Название телепередачи	Индекс Nielsen	Показания "пиплметров"
The Cosby Show	27,4	24,4
Family Ties	27,2	25,5
Cheers	24,1	23,0
Moonlighting	22,7	20,2
Night Court	22,1	20,1
Growing Pains	18,3	17,5
Who's the Boss	17,5	16,1
Family Ties (Special)	17,5	13,3
Murder She Wrote	16,8	14,7
60 Minutes	16,5	13,6

P. Barnes and J. Lipman, "Networks and Ad Agencies Battle over Estimates of TV Viewership", *The Wall Street Journal*, 1987, January 1, p. 25. В качестве источника использован NBC. Эти телешоу входили в десятку самых популярных в период с 10 по 23 октября. Один пункт рейтинга соответствует 1% от 97,7 миллиона телезрителей в возрасте от 25 до 54 лет.

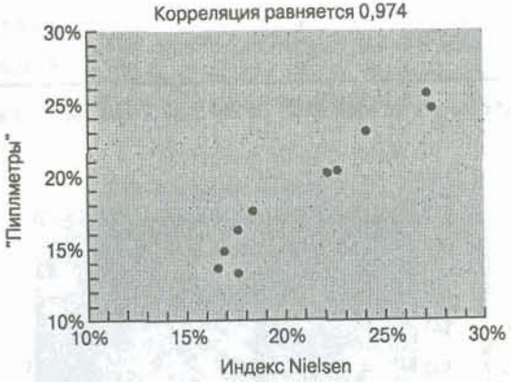


Рис. 11.1.3. Линейная взаимосвязь на диаграмме рассеяния для двух видов телевизионных рейтингов для n = 10 самых популярных телешоу. Обратите внимание на сильную положительную связь, результатом которой является высокая корреляция r = 0,974

Пример. Слияние компаний

Банкиры-инвесторы зарабатывают значительные суммы, предоставляя консультации и оказывая другие виды помощи компаниям, желающим объединиться или приобрести в собственность другие компании. Кто является крупнейшими участниками этой "игры"? Как много сделок и как много денег вовлечено в эту область деятельности? Ответы на эти вопросы можно найти, проанализировав двумерную совокупность данных из табл. 11.1.4, которая появилась примерно в то же время, когда некоторые из "звезд" вышли из состава First Boston, чтобы работать самостоятельно.

Таблица 11.1.4. Самые успешные фирмы, консультирующие по вопросам слияния и приобретения компаний в собственность

Название фирмы	Количество сделок	Сумма сделок, млн дол.
Goldman Sachs	134	63 465,0
First Boston	174	55 091,8
Morgan Stanley	120	42 336,3
Merrill Lynch	101	34 324,5
Shearson Lehman Brothers	164	25 631,7
Lazard Freres	44	24 251,5
Drexel Burnham Lambert	126	22 706,5
Salomon Brothers	76	21 859,7
Kidder Peabody	70	13 518,9
Dillon Read	42	11 167,8
Donaldson, Lufkin & Jenrette	47	7 750,1
Bankers Trust	41	5 525,7
PaineWebber	67	4 788,1
Allen & Co.	6	4 603,8
Bear Stearns	36	4 555,9

Данные взяты из "Top Advisers for Mergers and Acquisitions", *The Wall Street Journal*, 1988, February 3, p. 1. Источник данных: IDD Information Services.

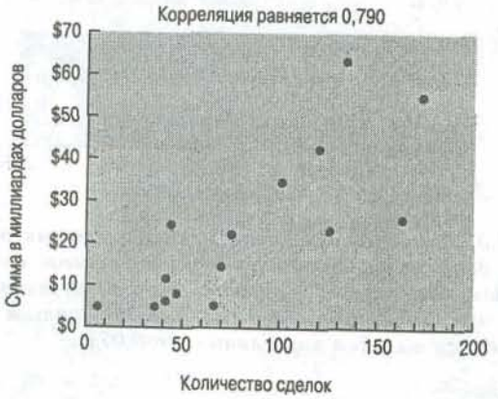


Рис. 11.1.4. Линейная взаимосвязь между суммой (в долларах) и количеством сделок, осуществляемых с участием крупнейших фирм, консультирующих по вопросам слияния и приобретения компаний в собственность. Корреляция r = 0,790 выражает сильную тенденцию роста (преуспевающие фирмы участвуют во множестве сделок с большими суммами денег), которая частично скрадывается действием фактора "случайности"

Диаграмма рассеяния, показанная на рис. 11.1.4, отражает линейную взаимосвязь, которая, однако, характеризуется значительно большей степенью разброса или случайности, чем в предыдущем примере. На диаграмме достаточно выражена тенденция роста, где более преуспевающие фирмы участвуют в большем количестве сделок (направление — вправо), в которых задействованы более крупные суммы денег (направление — вверх). Уже упоминавшаяся нами случайность касается довольно значительных денежных сумм: если, например, говорить о фирмах, участвующих примерно в 120 сделках, то величина этих сделок может различаться на десятки миллиардов долларов. Корреляция r= 0,790 отражает эту тенденцию к росту при наличии значительного фактора случайности.

Пример. *Ставки процента и комиссионные по закладной*

При получении денег под залог возникает множество разных расходов. Крупнейшими из них обычно являются *ссудный процент* (или ставка процента — годовой процент, который определяет размер вашего ежемесячного платежа) и *комиссионные за кредит* (одноразовая плата, которую с вас берут при предоставлении ссуды). Некоторые финансовые организации предлагают заемщику снизить ссудный процент, выплатив вначале повышенные комиссионные за кредит, предполагая при этом определенную взаимосвязь между этими двумя расходами. Эта взаимосвязь должна быть *отрицательной*, или *понижающей*, поскольку более высоким комиссионным за кредит должен соответствовать более низкий ссудный процент.

В табл. 11.1.5 приведена двумерная совокупность данных, включающая размеры ссудного процента и комиссионных за кредит для фирм, предоставляющих ссуды под залог недвижимости с фиксированным процентом сроком на 15 лет.

Диаграмма рассеяния, показанная на рис. 11.1.5, отражает линейную взаимосвязь, характеризующуюся значительным разбросом точек и *понижающей* связью между комиссионными за кредит и ссудным процентом. Отрицательная корреляция $r = -0,654$ подтверждает наличие предполагаемой нами понижающей взаимосвязи. О значительной доле случайности в этих данных свидетельствует то, что значение коэффициента корреляции находится примерно посередине между –1 и 0.

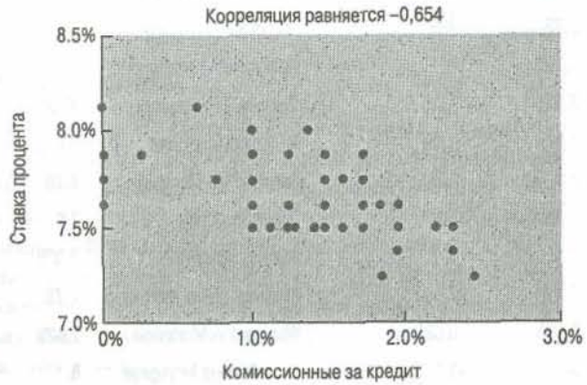


Рис. 11.1.5. Линейная взаимосвязь понижения между комиссионными за кредит и ссудным процентом для ссуды под залог недвижимости. Корреляция r = –0,654 отражает эту взаимосвязь понижения: более высоким комиссионным за кредит соответствуют более низкие значения ссудного процента. Поскольку значение коэффициента корреляции находится приблизительно посередине между –1 и 0, то данные характеризуются значительной случайностью

Таблица 11.1.5. Расходы, связанные с получением денег под залог недвижимости

Название организации	Ссудный процент, %	Комиссионные за кредит, %	Название организации	Ссудный процент, %	Комиссионные за кредит, %
Abacus Mortgage	7,25	1,875	Home Mortgage Corp.	7,625	1,5
Advocate Mortgage	7,875	1,5	Horizon Mortgage & Inv.	7,875	1
All American Mtg.	7,5	1,5	Intercontinental Mortgage	7,25	2,5
Alpine Mtg. Services	7,75	1	JE Mortgages Inc.	7,875	0
Alternative Mortgage	7,625	2	Madison Mortgage	8,125	0
Arboretum Mortgage	7,75	1	Mariner Mortgage, Inc.	7,625	2
Bancplus Mortgage	7,75	1,5	Mortgage Associates	7,75	0,75
Bancshares Mortgage Co.	7,5	1,625	Mortgage Network	7,5	1,75
Barigar Meier & Assoc.	7,5	1,125	Mortgage Solutions	7,875	1
Barto & Associates	7,875	1,5	Mortgage Brokers Service	7,5	1,42
Bay Mortgage	7,5	2	New World Mortgage	7,875	0,25
Best Mortgage Svs.	7,5	1,3	Normandy Mortgage	7,75	1
Bismark Mortgage	7,875	1	orth American Mortgage Co.	7,75	1,75
Capital Mortgage	7,75	1	Nu-West Mortgage	7,625	1
Carl I. Brown	7,625	1,875	Pacific Mountain Mortgage	7,375	2,375
Castle Mortgage Corp.	7,875	0	Principal Res. Mortgage	7,5	1,25
Chase Manhattan	7,625	1,75	Producer$ Mortgage Serv.	7,875	0
Concord Mortgage	7,75	1,5	Pro-West Fin. Group	7,5	2,25
Countrywide Funding	8,125	0	Qpoint Home Mortgage	7,5	2
Directors Mortgage	7,875	1,75	Raintree Fin. Network	7,625	0
Equity NW Inc.	7,875	1,25	Redmond Mortgage	7,875	0
First Am. Mtg. Group	7,5	1,5	Sammamish Mortgage	7,75	0
First Choice Financial	7,875	1,5	Select Mortgage	7,5	2,375
First Mark Mortgage	7,5	1	Sterling Mortgage	7,375	2
Fleet Mortgage Corp.	7,5	2	Stratford Home Mortgage	7.75	1,625
Group One Mortgage Inc.	8,125	0,625	Washington Mortgage	7,625	1,25
Guild Mortgage Co.	8	1,375	Wa. Womens Mortgage	8	1
Hallmark Mortgage	7,75	1	Western States Mortgage	7,5	1,5
Highland Res. Mortgage	7,875	0			

Данные получены из "Spring Mortgage Rates", *The Seattle Times*, 1995, April 23, p. G1. Источник данных: Scotsman Publishing, Inc.



Рис. 11.1.6. Предыдущая диаграмма рассеяния с добавлением эффекта "дрожания", позволяющего разделить перекрывающиеся точки и более отчетливо отобразить анализируемую совокупность данных

Куда же делись многие из этих данных? В двумерной таблице перечислено 57 финансовых организаций, однако создается впечатление, что количество точек на диаграмме рассеяния намного меньше этой величины. Это объясняется тем, что некоторые сочетания значений характеризуют несколько организаций (например, комиссионные за кредит, составляющие 1%, в сочетании со ссудным процентом, равным 7,75%). Эти несколько перекрывающихся точек выглядят на простой диаграмме (такой, которая показана на рис. 11.1.5) как одна точка. Добавив в диаграмму немного случайности, или "дрожания" (лишь для создания требуемого визуального эффекта, но не для анализа данных!), мы можем разделить эти перекрывающиеся точки и получить более отчетливое представление о соответствующих данных.[5] Полученная таким образом диаграмма разброса точек с "дрожанием" показана на рис. 11.1.6.

Отсутствие взаимосвязи

Взаимосвязь в двумерной совокупности данных **полностью отсутствует**, если соответствующая диаграмма рассеяния точек носит совершенно случайный характер, т.е. продвигаясь слева направо, мы не обнаруживаем тенденции направленности ни вверх, ни вниз. Случай полного отсутствия взаимосвязи представляет собой особый случай линейной взаимосвязи — без увеличения и уменьшения. Такая диаграмма рассеяния точек может иметь вид либо круглого, либо овального облака (причем овал может иметь вертикальную или горизонтальную ориентацию, однако не имеет наклона). Фактически, изменяя шкалу той или другой переменной, можно добиться того, что совокупность данных с полным отсутствием взаимосвязи будет иметь либо круговую, либо овальную диаграмму разброса точек.

Пример. "Инерция" и фондовая биржа

Обладает ли фондовая биржа какой-либо "инерцией"? Иными словами, должна ли цена акций сегодня расти только потому, что она росла вчера? Если существует какая-то взаимосвязь между поведением рынка вчера и его поведением сегодня, то можно надеяться на то, что эту взаимосвязь удастся выявить с

[5] Общее введение во множество различных методов визуализации данных (включая "дрожание", описанное на с. 135) приведено в книге Chambers J. M., Cleveland W. S., Kleiner B., and Tukey P. A. *Graphical Methods for Data Analysis* (New York: Wadsworth, 1983).

помощью соответствующей диаграммы рассеяния. В конце концов, это наш наилучший статистический инструмент, с помощью которого и можно выявить взаимосвязь (если она, разумеется, существует) между поведением рынка вчера (одна переменная) и его поведением сегодня (вторая переменная).

Двумерная совокупность данных включает суточную прибыль в соответствии с индексом S&P 500 Stock Market Index, т.е. процентные изменения (увеличение или уменьшение) индекса за день.[6] Несмотря на то что рассматриваемый нами случай напоминает одномерный временной ряд, одни и те же по сути данные можно записать в два столбца, сместив эти столбцы относительно друг друга на одну строку так, чтобы результат работы биржи за нынешний день (левый столбец в табл. 11.1.6) находился в той же строке, что и результат работы биржи за вчерашний день (правый столбец в табл. 11.1.6).

Диаграмма разброса точек, показанная на рис. 11.1.7, свидетельствует об отсутствии какой-либо взаимосвязи между двумя указанными переменными! Налицо совершенно произвольный разброс точек без какой-либо явно выраженной тенденции ориентации либо вверх (что указывало бы на наличие определенной "инерции" в поведении фондовой биржи), либо вниз (что указывало бы на то, что в какой-то день биржа проявила "чрезмерную реакцию", а затем просто "исправила" свое поведение) при перемещении по картинке слева направо. Корреляция $r = 0{,}11$ близка к 0, что свидетельствует об отсутствии сколько-нибудь значимой взаимосвязи.[7]



Рис. 11.1.7. Между нынешними и вчерашними результатами торгов на фондовой бирже нет никакой взаимосвязи. Корреляция r = 0,11 близка к 0, что указывает на отсутствие сильной взаимосвязи. Даже если вчера день на бирже был "хороший", нынешний день вполне может быть таким, будто вчера на бирже был "плохой" день

Диаграмма рассеяния, подобная приведенной на рис. 11.1.7, вполне соответствует теории эффективного рынка и теории "случайного блуждания". Теория эффективного рынка гласит, что вся имеющаяся информация или прогнозируемый ход событий немедленно отражаются на биржевых ценах. Поскольку трейдеры прогнозируют будущие изменения биржевых цен, говорить о каких-либо систематических взаимосвязях не приходится — остается одна лишь случайность (т.е. "случайное блуждание"). "Случайное блуждание" порождает временной ряд данных, в которых нет взаимосвязи между предшествующим поведением и последующим шагом, или изменением.[8]

Изменяя масштаб по вертикальной или горизонтальной оси, облаку точек можно придать вид, более похожий на линию. Однако поскольку такая линия будет либо *вертикальной*, либо *горизонтальной* — без какого-либо наклона, — это по-прежнему будет свидетельствовать об отсутствии взаимосвязи между переменными. Эти ситуации показаны на рис. 11.1.8 и 11.1.9.

Таблица 11.1.6. Процентное изменение индекса S&P 500 Stock Market Index

	Сегодня, %	Вчера, %		Сегодня, %	Вчера, %
1 мая 1998 г.	0,83	1,56		0,21	0,01
	0,10	0,83		–0,96	0,21
	–0,59	0,10		1,12	–0,96
	–0,95	–0,59		1,74	1,12
	–0,89	–0,95		0,17	1,74
	1,19	–0,89		0,24	0,17
	–0,14	1,19		–0,55	0,24
	0,83	–0,14		–1,59	–0,55
	0,28	0,83		0,39	–1,59
	–0,13	0,28		–1,99	0,39
	–0,77	–0,13		0,98	–1,99
	–0,26	–0,77		1,79	0,98
	0,33	–0,26		–0,07	1,79
	0,86	0,33		–0,52	–0,07
	–0,39	0,86		0,23	–0,52
	–0,37	–0,39		1,48	0,23
	–1,48	–0,37		1,20	1,48
	–0,16	–1,48		–0,32	1,20
	0,49	–0,16		0,35	–0,32
	–0,62	0,49		0,47	0,35
	0,01	–0,62	30 июня 1998 г.	–0,41	0,47

Нелинейная взаимосвязь

Теперь рассмотрим случай более сложного анализа двумерных совокупностей данных. В двумерной совокупности данных присутствует **нелинейная взаимосвязь**, если точки на диаграмме рассеяния группируются вокруг некоторой *кривой*, а не вокруг прямой линии. Поскольку разновидностей такого рода кривых может быть чрезвычайно много, анализ такой взаимосвязи существенно сложнее.

[6] Формула суточной прибыли имеет следующий вид: (нынешняя цена – вчерашняя цена)/(вчерашняя цена).

[7] Подобный этому коэффициент корреляции, вычисленный для временного ряда и его собственных предшествующих значений, называется *автокорреляцией* ряда, поскольку он определяет корреляцию этого ряда с самим собой. Можно сказать, что этот временной ряд не является сильно автокоррелированным, так как его коэффициент автокорреляции близок к нулю.

[8] Этому вопросу посвящена целая книга: Malkiel B. G. *A Random Walk down Wall Street* (New York: W. W. Norton, 1981).

Для совокупностей данных с нелинейной связью корреляционный и регрессионный анализ следует использовать с осторожностью. Применительно к некоторым задачам бывает полезно преобразовать одну или обе переменные таким образом, чтобы получить между ними линейную взаимосвязь. Это позволяет упростить анализ (поскольку корреляцию и регрессию удобнее применять именно к линейной взаимосвязи), а полученные результаты, если удается, преобразовывают обратно в исходную форму.[9]

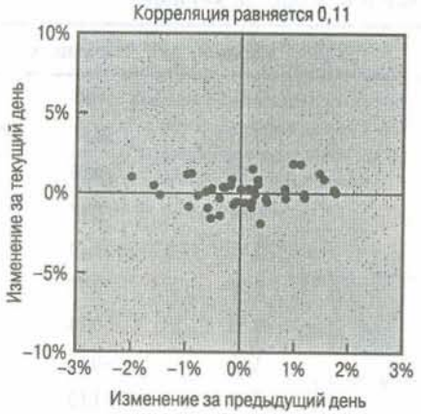


Рис. 11.1.8. В этом случае взаимосвязь между переменными также отсутствует, несмотря на то, что на диаграмме рассеяния явно просматривается прямая линия. Дело в том, что эта линия — горизонтальная, без наклона. На этой диаграмме представлена та же совокупность данных, что и на рис. 11.1.7, но уменьшен масштаб по оси Y, что сделало внешний вид диаграммы более плоским



Рис. 11.1.9. На этой диаграмме взаимосвязь между переменными также отсутствует, несмотря на то, что диаграмма рассеяния напоминает прямую линию. Дело в том, что эта линия — вертикальная, без наклона. На этой диаграмме уменьшен (по сравнению с рис. 11.1.7) масштаб по оси X

Пример. *Индексные опционы*

Если вы покупаете так называемый опцион *"колл"*, то получаете право — но не обязательство — купить какое-то имущество (которое может быть земельным участком, 100 акциями компании IBM и т.п.) по фиксированной цене (*цене использования опциона*) в любой момент, когда вам это понадобится, но лишь до тех пор, пока не истечет срок действия вашего опциона. Предприниматели пользуются опционами, чтобы подстраховаться от риска (т.е. снизить риск), заплатив за это существенно меньшую цену в сравнении с покупкой и, возможно, последующей продажей соответствующего имущества. Опционы на акции можно использовать либо для снижения риска определенного портфеля, либо для создания портфеля высокого риска с высоким ожидаемым доходом.

Чем выше цена использования опциона, тем меньше ценность этого опциона. Например, опцион на покупку плитки шоколада за $2000 бесполезен, но опцион на покупку этой же плитки за $0,25 имеет определенную ценность. Действительно, если цена плитки шоколада устойчива и равняется примерно $0,45, тогда ценность опциона составит $0,45 – $0,25 = $0,20. Однако для большинства рынков неопределенность относительно будущего повышает ценность опционов. Например, 19 января 1999 г. пакет акций Microsoft стоил $156, а опцион на покупку этого пакета акций в течение трех месяцев по цене $160 стоил примерно $12,00. Какой смысл был в том, чтобы покупать этот пакет за $160 на основе опциона, если его можно было купить за $156 прямо сейчас? Конечно, смысл во всем этом был бы лишь в том случае, если бы цена этого пакета поднялась до $170. Эта непредсказуемость поведения рынка (*возможность* повышения цен) частично объясняет ценность опциона.

Итак, мы рассчитываем обнаружить отрицательную взаимосвязь между ценой *использования опциона*, указываемой в контракте на опцион, и заявочной ценой, по которой продается сам контракт на опцион. В табл. 11.1.7 представлена двумерная совокупность данных для самых популярных индексных опционов, основанная на индексе Standard & Poor's 100.

Диаграмма рассеяния, представленная на рис. 11.1.10, иллюстрирует пример нелинейной взаимосвязи. Эта взаимосвязь носит отчетливо выраженный отрицательный характер, поскольку чем выше цена использования опциона, тем ниже заявочная цена. Корреляция $r = -0,895$ служит подтверждением сильной отрицательной взаимосвязи. Поскольку эта взаимосвязь почти идеальна, а элемент случайности практически полностью отсутствует, можно было бы ожидать, что коэффициент корреляции будет еще ближе к –1. Однако это могло бы произойти лишь в том случае, если бы точки располагались строго на *прямой линии*. Поскольку же точки располагаются строго на кривой линии, корреляция отличается от –1.

Более совершенные статистические методы, основанные на предположении о нормальном распределении и случайном блуждании цен акций, дали возможность аналитикам вычислить приблизительное значение цены для опциона "колл".[10] Эта сложная и продвинутая теория основана на тщательном вычислении математического ожидания (среднего значения) случайной переменной, представляющей максимальную плату за соответствующий опцион.

Таблица 11.1.7. Опционы "колл" для индекса S&P 100

Цена использования опциона, дол.	Заявочная цена, дол.	Цена использования опциона, дол.	Заявочная цена, дол.
470	80	565	5
510	40,375	570	3,375
515	35,625	575	3
520	33,875	580	1,75
525	27,875	585	1,125
530	23,75	590	0,875
535	21	595	0,5
540	17	600	0,375
545	14	605	0,1875
550	11	620	0,125
555	9	630	0,125
560	6,5		

Данные об опционах, срок действия которых истекает в сентябре. Заимствовано из *The Wall Street Journal*, 1998, August 20, p. C14.

[9] Использование преобразований в регрессии будет рассмотрено в главе 12.

[10] Обзор теории и практики опционов приведен в книге Cox J. C. and Rubenstein M. *Options Markets* (Englewood Cliffs, N.J.: Prentice Hall, 1985).



Рис. 11.1.10. Нелинейная взаимосвязь между ценой опциона и ценой использования опциона. На диаграмме прослеживается вполне ожидаемая отрицательная взаимосвязь — но нелинейная, *поскольку линия, образуемая точками, оказывается кривой. Корреляция, r = –0,895, свидетельствует о сильной отрицательной взаимосвязи. Из-за криволинейности коэффициент корреляции не может равняться в точности –1, несмотря на то, что в данном случае мы имеем дело с практически идеальной взаимосвязью (случайного разброса точек почти нет)*

Пример. Объем выпускаемой продукции и температура

Сильная нелинейная взаимосвязь может быть даже тогда, когда корреляция близка к нулю! Это может произойти в случае, если эта сильная взаимосвязь не является ни увеличивающейся, ни уменьшающейся (что бывает при наличии оптимального, или наилучшего из возможных значений). Рассмотрим данные, полученные в результате эксперимента, целью которого являлось определение такой температуры, которая обеспечивает для определенного промышленного процесса максимальный объем выпуска продукции. Соответствующие данные приведены в табл. 11.1.8.

Диаграмма рассеяния, показанная на рис. 11.1.11, иллюстрирует сильную нелинейную взаимосвязь, характеризующуюся незначительным случайным разбросом. Коэффициент корреляции, $r = –0,0155$, бесполезен в случае такой нелинейной связи: он не может решить, является связь увеличивающей или уменьшающей, поскольку в действительности есть и то и другое.

В этом случае диаграмма рассеяния очень полезна, поскольку демонстрирует, что для максимального увеличения объема выпускаемой продукции температуру производственного процесса следует установить равной примерно 700 градусам. Объем продукции резко падает как при слишком высокой, так и при слишком низкой температуре. Этот важный вывод можно сделать, наблюдая на диаграмме сильную взаимосвязь между объемом продукции и температурой.

Помните: близкое к нулю значение корреляции может означать как отсутствие взаимосвязи в данных, так и наличие нелинейной взаимосвязи без преобладания направленности вниз или вверх.

Неодинаковая вариация

Еще одна техническая трудность, которая, к сожалению, нередко встречается в данных, касающихся бизнеса и экономики, заключается в том, что изменчивость (вариация) по вертикальной оси на диаграмме рассеяния может зависеть от

Таблица 11.1.8. Температура и объем продукции для промышленного процесса

Температура, градусы	Объем продукции	Температура, градусы	Объем продукции
600	127	750	153
625	139	775	148
650	147	800	146
675	147	825	136
700	155	850	129
725	154		

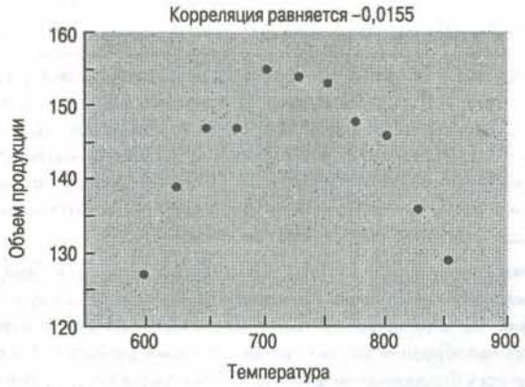


Рис. 11.1.11. Нелинейная взаимосвязь между объемом продукции и температурой для некоторого промышленного процесса. Существует сильная взаимосвязь, но она нелинейная. Коэффициент корреляции, r = –0,0155, свидетельствует лишь о том, что в целом *нет определенной направленности — ни вверх, ни вниз*

того, где вы находитесь в данный момент на горизонтальной оси. Когда речь идет об анализе деятельности крупных компаний (или других элементарных единиц анализа), обнаруживается весьма значительная изменчивость, величина которой измеряется, возможно, миллионами или даже миллиардами долларов, но когда вы анализируете деятельность небольших предприятий, изменчивость может измеряться величинами порядка десятков тысяч долларов. Считается, что у диаграммы рассеяния **неравная вариация**, если при перемещении по горизонтальной оси диаграммы рассеяния величина вариации по вертикальной оси изменяется очень сильно.[11]

Проблема неодинаковой вариации заключается в том, что места, характеризующиеся высокой изменчивостью, представляют *наименее точную* информацию и в то же время, как правило, оказывают наибольшее влияние на статистические показатели. Поэтому, если вы получили диаграмму разброса точек с чрез-

[11] Технические термины *гетероскедастичный* (прилагательное) и *гетероскедастичность* (существительное) также описывают неодинаковую вариацию.

вычайно неодинаковой изменчивостью, соответствующий коэффициент корреляции (и другие характеристики такой взаимосвязи) будет ненадежным.

Эту проблему зачастую удается решить путем преобразования данных — возможно, с помощью логарифмов. К счастью, такое преобразование, если его применить к каждой переменной, нередко позволяет решить сразу несколько проблем. Во многих случаях удается не только выровнять изменчивость, но и приблизить к нормальному распределения самих переменных. Логарифмы (можно использовать как натуральные по основанию *e*, так и привычные по основанию 10), как правило, очень хорошо подходят для работы с денежными суммами. Преобразование с помощью извлечения квадратного корня хорошо подходит для работы с количеством каких-либо вещей или событий.

Пример. *Оптический кабель*

Многие инвестиции в сферу высоких технологий являются достаточно рискованными: применяемые в них методы новы и еще неизвестно, как они себя проявят на практике, дорогостоящее оборудование относительно быстро устаревает, а конкуренция в этой сфере бывает весьма жесткой. Тем не менее миллиарды долларов инвестируются в создание магистральных коммуникационных сетей на основе оптических кабелей. Если вас интересует этот вид инвестиций, вам, наверное, интересно было бы узнать также, во сколько обходится создание таких оптоволоконных сетей связи. Соответствующая информация о ведущих компаниях, работающих в этой области, приведена в табл. 11.1.9.

Диаграмма рассеяния, показанная на рис. 11.1.12, иллюстрирует в целом взаимосвязь увеличивающегося типа: фирмы, осуществляющие более крупные капиталовложения, как правило, создают сети связи с более высокой общей протяженностью. В то же время изменчивость характеризуется значительным неравенством, о чем свидетельствует воронкообразная форма соответствующих данных (с "раструбом", направленным вправо). Данные, относящиеся к более мелким фирмам, группируются внизу и слева, указывая на весьма не-

Таблица 11.1.9. Оптоволоконные магистральные сети связи

	Капиталовложения, млн дол.	Сетевые мили*, млн дол.
AT&T	1300	1700
MCI	500	650
GTE	130	110
United Telecommunications	2000	1200
Fibertrak	1200	2400
LDX Net	110	165
Electra Communications	40	72
Microtel	60	45
Litel Telecommunications	57	85
Lightnet	500	650
SoutherNet	90	50
RCI	90	87

* *Сетевая миля* определяется как протяженность кабеля, способного передавать один речевой сигнал на расстояние в одну милю.

Данные получены из статьи W. B. Johnston, "The Coming Glut of Phone Lines", *Fortune*, 1985, January 7, p. 97–100. Источник данных: Hudson Institute.

значительную изменчивость в общей протяженности создаваемых ими сетей; у более крупных фирм (справа) проявляется намного большая изменчивость в общей протяженности сетей, создаваемых ими на основе своих более крупных инвестиций. Рис. 11.1.13 демонстрирует, какие именно изменчивости являются неравными: изменчивости измерений по вертикали и касающиеся протяженности сетей.

Может ли преобразование помочь решить эту проблему неодинаковой изменчивости? Попробуем, например, применить натуральные логарифмы. Капиталовложения AT&T, например, составляют $1 300 000 000; соответствующий логарифм равен 21,0. Общая протяженность сетей у AT&T равняется 1 700 000 000; логарифм этого числа — 21,3. Результаты вычисления логарифмов всех значений приведены в табл. 11.1.10.

Диаграмма разброса точек, показанная на рис. 11.1.14, иллюстрирует неплохую линейную взаимосвязь между логарифмом суммы капиталовложений и логарифмом общей протяженности сетей. Эта диаграмма выглядела бы точно так же, если бы мы использовали обычные логарифмы (по основанию 10) или если бы мы вместо реальных значений (например, $1 300 000 000) логарифмировали значения, выраженные в миллионах (например, $1 300). Таким образом, нам удалось избавиться от неодинаковой вариации.

Как правило (хоть и не всегда), после использования подходящего преобразования коэффициент корреляции несколько возрастает: в данном случае с 0,820 (до преобразования) до 0,957 (после логарифмического преобразования). Таким образом, в целом можно утверждать, что, поскольку корреляция в исходном масштабе так сильно зависит от небольшого количества очень крупных фирм, коэффициент корреляции после преобразования данных является более надежным показателем связи.



Рис. 11.1.12. Неодинаковая изменчивость во взаимосвязи между общей протяженностью линий и капиталовложениями. У крупных компаний (справа) наблюдается намного большая изменчивость, чем у небольших (слева)

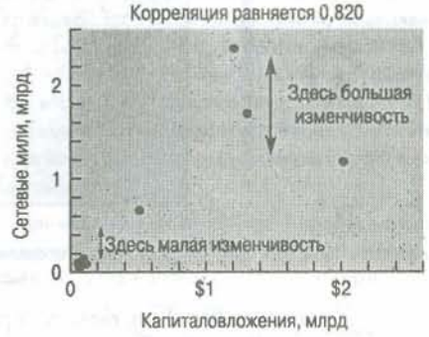


Рис. 11.1.13. Диаграмма рассеяния, определяющая взаимосвязь между общей протяженностью линий и капиталовложениями. Отчетливо обозначена неодинаковая изменчивость у крупных и небольших компаний

Таблица 11.1.10. Оптоволоконные магистральные сети связи

	Капиталовложения (логарифм денежных сумм)	Сетевые мили (логарифм)
AT&T	21,0	21,3
MCI	20,0	20,3
GTE	18,7	18,5
United Telecommunications	21,4	20,9
Fibertrak	20,9	21,6

Окончание табл. 11.1.10

	Капиталовложения (логарифм денежных сумм)	Сетевые мили (логарифм)
LDX Net	18,5	18,9
Electra Communications	17,5	18,1
Microtel	17,9	17,6
Litel Telecommunications	17,9	18,3
Lightnet	20,0	20,3
SoutherNet	18,3	17,7
RCI	18,3	18,3

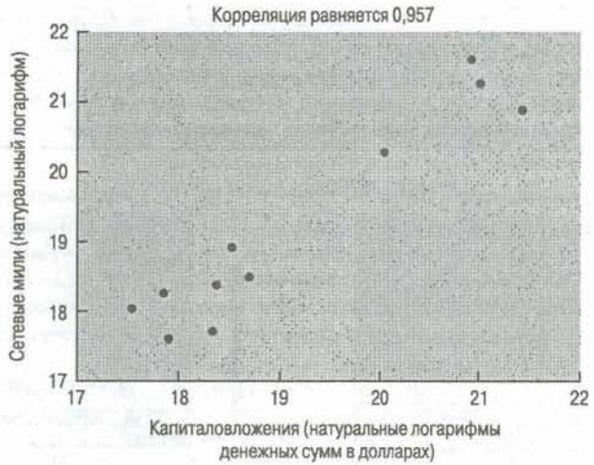


Рис. 11.1.14. Результат преобразования в линейную взаимосвязь между натуральными логарифмами общей протяженности линий и объемами капиталовложений. Воспользовавшись подобным преобразованием, можно решить проблему неравной изменчивости. Эта совокупность данных — в логарифмическом масштабе — характеризуется линейной взаимосвязью

Разделение совокупности на группы

Говорят, что в двумерной совокупности данных наблюдается **разделение на группы (кластеринг)**, если на соответствующей диаграмме рассеяния видны отдельные, отличные одна от другой группы точек (кластеры). Если в данных действительно есть отдельные группы, но вы этого не осознаете, у вас могут быть серьезные проблемы, поскольку обычные статистические показатели взаимосвязи не могут учитывать такой вид взаимосвязи. Ваша задача заключается в том, чтобы выявить кластеринг в своих данных и (если он имеется) предпринять надлежащие меры, разделив, например, всю совокупность данных на две или несколько частей, каждая из которых соответствует отдельному кластеру.

Типичная проблема, которая возникает в случае кластеринга, заключается в том, что в каждом кластере существует четкая взаимосвязь, но коэффициент корреляции для всей совокупности данных указывает на отсутствие взаимосвязи. Хуже того, коэффициент корреляции может показывать взаимосвязь во всей совокупности данных, *совершенно противоположную* взаимосвязи в каждом отдельном кластере! Всегда старайтесь анализировать каждую диаграмму рассеяния именно с точки зрения возможного наличия кластеров: сама по себе корреляция еще ни о чем не говорит.

Пример. "Цветочные" облигации

Облигации Казначейства США принадлежат к числу наименее рисковых инвестиций — с точки зрения вероятности того, что вы действительно получите все причитающиеся вам выплаты.[12] Помимо проводимой Казначейством публичной продажи облигаций на первичном рынке ценных бумаг, существует активный вторичный рынок, на котором выставляются на продажу все нереализованные ценные бумаги. Нетрудно предположить существование взаимосвязи увеличивающего типа между купоном облигации, который указывает величину периодических выплат по этой облигации (дважды в год), и текущей ценой, по которой продается облигация. В табл. 11.1.11 показана двумерная совокупность данных: учетная ставка купона и цена покупателя (цена, или курс, по которой покупатель согласен приобрести ценную бумагу) для облигаций Казначейства США, срок погашения которых наступает в период с 1994 по 1998 гг.

На диаграмме разброса точек, показанной на рис. 11.1.15, отчетливо виден кластеринг. Обычные облигации образуют один кластер с очень сильной линейной взаимосвязью. Тщательный анализ позволяет установить, что три особые облигации (долговые обязательства), образующие кластер в нижнем левом углу диаграммы, относятся к типу так называемых *"цветочных" облигаций*. Эти три "цветочные" облигации образуют кластер с особой взаимосвязью между купоном и ценой. Суммарная корреляция, $r = 0{,}867$, указывает на силу взаимосвязи между всеми точками данных во всех кластерах. Взаимосвязь между обычными облигациями оказывается намного сильнее и характеризуется коэффициентом корреляции $r = 0{,}993$, который вычисляется без учета трех цветочных облигаций.

Что произошло бы, если бы мы не обратили внимание на наличие кластеров? Можно было бы, например, прийти к ошибочному заключению, что взаимосвязь между купоном и ценой просто "достаточно сильная" и характеризуется коэффициентом корреляции $r = 0{,}867$, хотя на самом деле взаимосвязь между обычными облигациями оказывается намного сильнее (ее можно охарактеризовать как "очень сильную" с коэффициентом корреляции $r = 0{,}993$). Если эту совокупность данных использовать для определения цен или для принятия решения, какие из обычных облигаций следует продать, полученные результаты были бы искажены наличием "цветочных" облигаций. В определенном смысле цветочные облигации представляют собой иную разновидность защиты гарантии от риска, и было бы неправильным указывать их в одних списках с другими типами облигаций.

Что же такое "цветочные" облигации и почему для них характерен столь большой разброс цен? От облигаций других типов их отличает то, что с их помощью решается вопрос уплаты налогов. Эти облигации погашаются по номиналу (т.е. по их номинальной стоимости) в порядке уплаты налогов на наследство. Если вы очень богатый человек и чувствуете приближение смерти (что повлечет за собой необходимость платить налоги на наследство), вам, может быть, имеет смысл приобрести по $94 облигации, номинальная стоимость которых равняется $100. Всем остальным вряд ли стоит задумываться о покупке таких облигаций. В соответствии с диаграммой рассеяния (смотрим на взаимосвязь для обычных облигаций) по причине низких купонных выплат эти облигации должны стоить не более $80. Однако ценность этих облигаций как средства уплаты налогов на наследство существенно поднимает их цену.

[12] Если, однако, вы примете решение продать облигацию до указанного срока ее погашения, вам придется столкнуться с так называемым "процентным риском", т.е. риском потерь в результате изменения процентных ставок (например, стоимость облигации с фиксированной ставкой уменьшается по мере общего повышения процентных ставок).

Выбросы (резко отклоняющиеся значения)

Точка данных на диаграмме рассеяния представляет собой **выброс (резко отклоняющееся значение)**, если она не соответствует взаимосвязи, присущей остальным данным. Выбросы могут так искажать статистические характеристики, что лишь вводят нас в заблуждение. Анализируя диаграмму рассеяния, вы всегда должны проверять данные на наличие выбросов. Если можно оправдать удаление таких резко отклоняющихся значений (полагая, например, что их появление носит исключительно случайный характер), от них следует избавиться. Если вам все же придется оставить их, то по крайней мере следует помнить о проблемах, которые эти точки могут порождать, и рассмотреть возможность получения статистических характеристик (например, коэффициента корреляции) как с учетом, так и без учета этих резко отклоняющихся значений.

Выброс может искажать корреляцию, которая в некоторых случаях может указывать на сильную взаимосвязь, тогда как на самом деле ничего, кроме случайности и одного резко отклоняющегося значения, нет. Выброс может также искажать корреляцию таким образом, что создается впечатление *отсутствия*

Таблица 11.1.11. Облигации Казначейства США

Учетная ставка купона, %	Цена покупателя, дол.	Учетная ставка купона, %	Цена покупателя, дол.
7,000	92,94	12,625	119,06
9,000	101,44	8,875	100,38
7,000	92,66	10,500	108,50
4,125	94,50	8,625	99,25
13,125	118,94	9,500	103,63
8,000	96,75	11,500	114,03
8,750	100,88	8,875	100,38
12,625	117,25	7,375	92,06
9,500	103,34	7,250	90,88
10,125	106,25	8,625	98,41
11,625	113,19	8,500	97,75
8,625	99,44	8,875	99,88
3,000	94,50	8,125	95,16
10,500	108,31	9,000	100,66
11,250	111,69	9,250	102,31
8,375	98,09	7,000	88,00
10,375	107,91	3,500	94,53
11,250	111,97		

Данные получены из *The Wall Street Journal*, 1988, November 9, p. C19. Цены покупателя указаны из расчета на "номинальную стоимость", равную $100 и выплачиваемую при погашении облигации. Половина купона оплачивается каждые шесть месяцев. Например, для первого элемента таблицы каждые шесть месяцев до наступления срока погашения выплачивается $3,50 (половина 7% купона); при погашении облигации выплачивается еще $100.

взаимосвязи, когда фактически имеется сильная взаимосвязь и одно резко отклоняющееся значение. Можно ли защититься от подобных ловушек? Можно, для этого нужно лишь внимательно проанализировать диаграмму рассеяния.



Рис. 11.1.15. Кластеринг во взаимосвязи между ценой покупателя и выплатой по купонам облигаций. Цены обычных облигаций отличаются от цен "цветочных" облигаций, поэтому каждый из кластеров характеризуется собственной взаимосвязью. Суммарная корреляция, $r = 0,867$, не учитывает взаимосвязи внутри каждого отдельного кластера. Корреляция для кластера обычных облигаций, $r = 0,993$, оказывается намного выше

Пример. Количество произведенных изделий и затраты

Рассмотрим количество изделий, выпускаемых каждую неделю на заводе, и общие затраты за ту же неделю. В данном случае должна наблюдаться достаточно сильная взаимосвязь между переменными. В те недели, когда завод работает с полной нагрузкой, выпускается большое количество изделий, для производства которых требуются большие объемы исходных материалов, — соответственно возрастают и затраты. Однако в имеющихся данных нас ожидают некоторые сюрпризы. В данных, представленных в табл. 11.1.12, имеется отрицательная корреляция $r = -0,623$. Почему корреляция отрицательна?

Диаграмма рассеяния, показанная на рис. 11.1.16, содержит одно резко отклоняющееся значение. Этим и объясняется отрицательная корреляция — даже несмотря на то, что остальные данные демонстрируют определенную положительную связь (которая, однако, нивелируется имеющимся выбросом). Это лишний раз подтверждает наш вывод о том, что резко отклоняющиеся значения необходимо тщательно анализировать. В данном случае наличие такого значения объясняется пожаром, который произошел на заводе. Значительная часть исходных материалов была уничтожена огнем — именно это и стало причиной резкого скачка затрат на той неделе. Выпуск продукции резко упал из-за остановки производства в 11.00, и даже не всю продукцию, соответствующую этой точке, можно было использовать.

Имеем ли мы право игнорировать это резко отклоняющееся значение? В данном случае, видимо, имеем. Несомненно, имеем, если нас интересует взаимосвязь в обычные недели и если пожар рассматривается как исключительный случай, нетипичный для нормальных обстоятельств. Действительно, если мы отбросим это резко отклоняющееся значение, корреляция станет положительной и близкой к 1 величиной ($r = 0,869$), указывая на достаточно сильную взаимосвязь увеличивающего типа между затратами и объемом выпускаемой продукции.

Совокупность данных без выброса показана на рис. 11.1.17. Обратите внимание: отбросив резко отклоняющееся значение, можно увеличить масштаб диаграммы и более подробно рассмотреть оставшиеся данные.

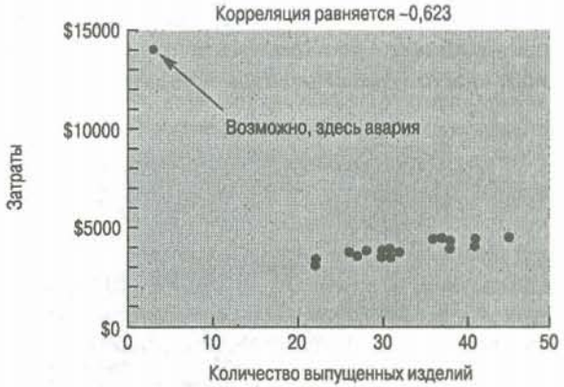


Рис. 11.1.16. Резко отклоняющееся значение нарушило корреляцию. Вместо того чтобы выявить в целом взаимосвязь роста между объемом произведенной продукции и затратами, коэффициент корреляции, r = –0,623, указывает на наличие взаимосвязи уменьшения, при которой более высоким объемам производства соответствуют меньшие затраты

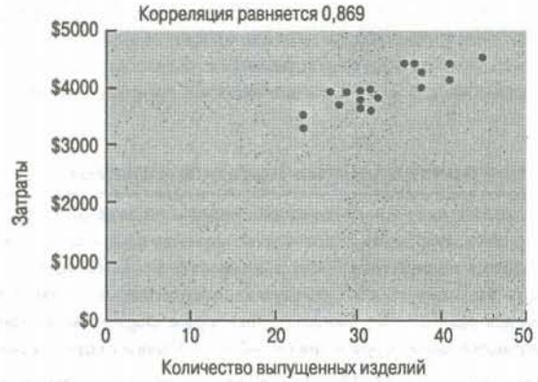


Рис. 11.1.17. Та же совокупность данных, но без резко отклоняющегося значения, иллюстрирует взаимосвязь между объемом произведенной продукции и затратами для "обычных" недель (без чрезвычайных происшествий). Коэффициент корреляции, r = 0,869, в этом случае имеет положительное значение и приближается к 1, что указывает на взаимосвязь роста

Корреляция — это не причинная обусловленность

Очень часто корреляцию и причинную обусловленность считают синонимами. Такое представление имеет под собой определенные основания, поскольку, когда нечто является *причиной* чего-либо другого, можно говорить о связи первого и второго и, следовательно, об их коррелированности (например, действие и результат, проверка и качество, капиталовложения и прибыль, окружающая среда и производительность).

Таблица 11.1.12. Недельный объем производства

Количество изделий	Затраты, дол.	Количество изделий	Затраты, дол.
22	3470	30	3589
30	3783	38	3999
26	3856	41	4158
31	3910	27	3666
36	4489	28	3885
30	3876	31	3574
22	3221	37	4495
45	4579	32	3814
38	4325	41	4430
3	14 131		

Однако корреляция бывает и без причинной обусловленности. Это можно представить себе так: корреляция — лишь число, которое указывает на то, что большим значениям одной переменной соответствуют большие (или, наоборот, малые) значения другой переменной. Корреляция не может объяснить, *почему* эти две переменные связаны между собой. Действительно, корреляция никак не объясняет, почему капиталовложения порождают прибыль (или наоборот)! Корреляция просто указывает, что между этими величинами наблюдается определенное соответствие.

Одним из возможных оснований для существования "корреляции без причинной обусловленности" является наличие некоторого скрытого, ненаблюдаемого, *третьего фактора*, создающего *впечатление*, будто одна переменная является причиной другой переменной, тогда как на самом деле причиной для каждой из этих двух переменных является эта неизвестная третья переменная. Термином **ложная корреляция** обозначают высокую корреляцию, которая на самом деле обеспечивается действием некоего "третьего фактора". Допустим, вы обнаружили высокую корреляцию между приемом на работу новых менеджеров и созданием новых производственных мощностей. Может быть, именно новые менеджеры являются "причиной" капиталовложений в новые производственные мощности? Или, наоборот, создание новых производственных мощностей послужило "причиной" приема на работу новых менеджеров? Скорее всего, однако, здесь проявляется действие третьего фактора: высокой, рассчитанной на длительную перспективу потребности в продукции фирмы, которая и послужила причиной и приема на работу новых менеджеров, и создания новых производственных мощностей.

Пример. *Расходы в продовольственных магазинах и ресторанах*

Основываясь на данных из различных штатов, показанных на диаграмме рассеяния (рис. 11.1.18), нетрудно прийти к выводу об очень высокой корреляции ($r = 0,988$) между суммой денежных расходов в продовольственных магазинах и ресторанах (в местах, где можно поесть и выпить), причем эта корреля-

ция имеет высокий уровень значимости ($p < 0{,}001$).[13] Чтобы разобраться в этом, постараемся прежде всего ответить на следующий вопрос: "Является ли привычка тратить много денег в продовольственных магазинах "причиной", которая заставляет человека тратить много денег в ресторанах?" Лично я так не считаю. Что касается меня, то чем больше денег я трачу в продовольственных магазинах, тем реже посещаю рестораны: действительно, зачем мне идти в ресторан, если у меня хватает еды дома? Может быть, здесь причинно-следственная связь имеет другую направленность: "Привычка тратить много денег в ресторанах является "причиной", которая заставляет человека тратить много денег в продовольственных магазинах?" Однако рассуждения, подобные приведенным выше, заставляют нас и в этом случае дать отрицательный ответ, поскольку человеку, который тратит много денег в ресторанах, скорее всего нет нужды хранить у себя дома большой запас продуктов. Вообще говоря, экономисты считают, что рестораны и продовольственные магазины в какой-то степени заменят друг друга.

Если ни одна из переменных (расходы в продовольственных магазинах и расходы в ресторанах) не является непосредственной причиной изменения другой переменной, тогда, может быть, существует некий третий фактор, влияющий на обе эти переменные? Может быть, этим третьим фактором является численность населения штата?[14] Соответствующие коэффициенты корреляции весьма высоки: $r = 0{,}994$ (между численностью населения и расходами в продовольственных магазинах) и $r = 0{,}990$ (между численностью населения и расходами в ресторанах). Весьма правдоподобным является такое объяснение: в штатах с большим населением тратят денег больше *и* в продовольственных магазинах, *и* в ресторанах, просто поскольку в этих штатах проживает больше людей! Связь между расходами в продовольственных магазинах и ресторанах является косвенной и объясняется достаточно просто наличием указанного третьего фактора.[15]

11.2. Регрессия: предсказание одного фактора на основании другого

Регрессионный анализ позволяет предсказывать одну переменную на основании другой с использованием прямой линии, характеризующей взаимосвязь между этими двумя переменными. Переменную, поведение которой прогнозируется, принято обозначать буквой Y; переменную, которая используется для такого прогнозирования, принято обозначать буквой X. Очень важно, что вы определяете как X и Y, поскольку *X предсказывает Y*, и *Y предсказывается с помощью X*. В табл. 11.2.1 представлен ряд стандартных способов описания роли каждой из переменных и соответствующие примеры.

[13] Основано на данных за 1995 г. о 50 штатах и Округе Колумбия, приведенных в таблице 1277 Бюро переписи населения США, *Statistical Abstract of the United States: 1997*, 117th edition (Washington, D.C.) 1997. Описание проверки значимости для двумерных данных будет рассмотрено в разделе 11.2.

[14] Основано на данных переписи населения за 1996 г., приведенных в таблице 33 Бюро переписи населения США, *Statistical Abstract of the United States: 1997*, 117th edition (Washington, D.C.) 1997.

[15] В действительности с помощью остаточного анализа можно убедиться в отсутствии значимой взаимосвязи между изменением расходов в продовольственных магазинах и ресторанах после учета численности населения штатов. Это можно сделать, определив остатки от регрессии для прогнозирования расходов в продовольственных магазинах на основе численности населения (эти остатки показывают, насколько больше — или, наоборот, меньше — оказываются расходы в данном штате по сравнению с тем, что можно было бы ожидать от штата с такой численностью населения) и вычислив их связь с остатками от регрессии для прогнозирования расходов в ресторанах на основе численности населения. В главе 12, посвященной множественной регрессии, будет описан другой метод учета дополнительных переменных.

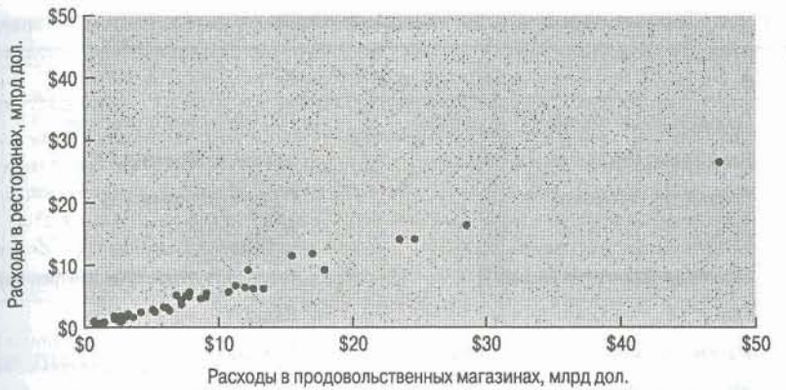


Рис. 11.1.18. Корреляция без непосредственной причинно-следственной связи между расходами в продовольственных магазинах и ресторанах в различных штатах. Наблюдается очень сильная положительная взаимосвязь ($r = 0{,}988$), свидетельствующая о том, что высоким расходам в ресторанах соответствуют высокие расходы в продовольственных магазинах, несмотря на то, что эти способы питания в известной степени являются экономически взаимозаменяемыми и что люди, часто посещающие рестораны, должны были бы тратить меньше *денег на походы в продовольственные магазины. Высокие расходы в продовольственных магазинах не являются непосредственной "причиной" высоких расходов на питание в ресторанах; на взаимосвязь между этими расходами косвенно влияют различия в численности населения штатов: в штатах с большей численностью населения, как правило, отмечаются более высокие расходы как на питание в ресторанах, так и на покупки в продовольственных магазинах*

Прямая линия характеризует линейную взаимосвязь

Термином **линейный регрессионный анализ** обозначают прогнозирование одной переменной на основании другой, когда между этими переменными существует линейная взаимосвязь. Точно так же как понятие "среднего значения" можно использовать в качестве характеристики отдельной переменной, прямая линия может выступать в качестве характеристики предполагаемой линейной связи между двумя переменными. Точно так же как для одномерных данных существует изменчивость относительно среднего значения, для двумерных данных существует изменчивость относительно соответствующей прямой линии. Так же как среднее значение, прямая линия является весьма полезной, но все же не идеальной характеристикой, поскольку присутствует случайность.

На рис. 11.2.1 показана прямая линия, характеризующая данные телевизионных рейтингов, — пример линейной взаимосвязи, рассмотренный нами ранее в этой главе. Обратите внимание, как эта линия отражает взаимосвязь возрастания. Она отражает основную структуру в этих данных: точки на диаграмме лишь случайно отклоняются от этой прямой линии (случайные флуктуации).

После краткого обсуждения таких прямых линий мы покажем, как вычислять и интерпретировать линию регрессии, как вычислить, насколько эта линия хорошо соответствует данным, как исходя из выборки делать правильный вывод относительно генеральной совокупности и как учитывать возможные сложности.

Таблица 11.2.1. Переменные в регрессионном анализе

	X	Y
Роли	Прогнозирующая переменная (предиктор)	Прогнозируемая переменная
	Независимая переменная	Зависимая переменная
	Поясняющая переменная	Поясняемая переменная
	Стимул	Реакция
	Экзогенная переменная (внешняя)	Эндогенная переменная (внутренняя)
Примеры	Сбыт	Доходы
	Количество произведенной продукции	Затраты
	Затраченные усилия	Полученные результаты
	Капиталовложения	Выпуск продукции
	Практический опыт	Заработная плата
	Температура процесса	Объем произведенной продукции

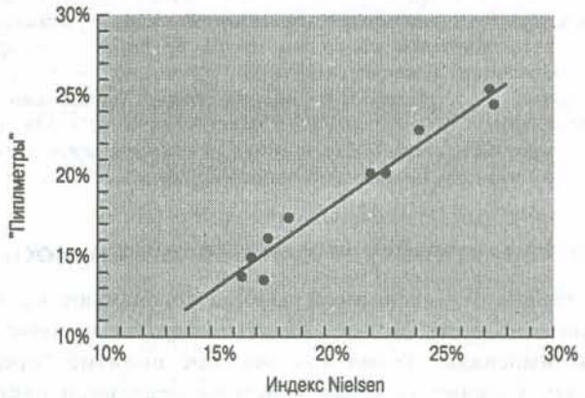
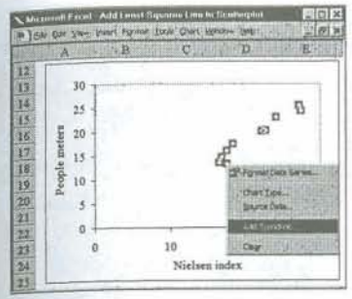
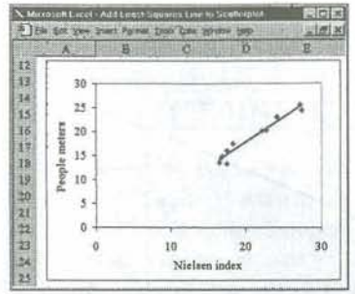


Рис. 11.2.1. Линия регрессии характеризует взаимосвязь между двумя оценками величины телевизионной зрительской аудитории. Эта линия показывает, как можно прогнозировать рейтинги, полученные с помощью "пиплметров" (Y), на основании индекса Nielsen (X)

Чтобы с помощью Excel® нанести на диаграмму рассеяния линию наименьших квадратов, достаточно щелкнуть правой кнопкой мыши *на какой-либо точке данных* на диаграмме, выбрать в появившемся на экране контекстно-зависимом меню команду Add Trendline (Добавить линию тренда) и, наконец, прежде чем щелкнуть на кнопке OK, выбрать в качестве типа линейную регрессию. Ниже показан начальный шаг (щелчок правой кнопкой мыши на точке данных), за которым следует результат, полученный после добавления линии.





Прямые линии

Прямая линия описывается двумя значениями: *наклоном*, b, и *сдвигом*, a. **Наклон** указывает на крутизну подъема (или снижения — если значение b отрицательно) линии. Если сместиться по горизонтали вправо ровно на 1 единицу измерения X, линия поднимется (или снизится, если $b<0$) по вертикали на b единиц измерения Y. **Сдвиг** — это просто значение Y, когда X равно 0. В случаях, когда нулевое значение X лишено смысла, сдвиг следует рассматривать как технически необходимую характеристику линии и его не следует непосредственно интерпретировать.[16] Уравнение прямой линии имеет следующий вид.

Уравнение прямой линии

$$Y = \text{Сдвиг} + (\text{Наклон})(X) = a + bX$$

Наклон и сдвиг показаны на рис. 11.2.2–11.2.4.

Построение линии на основе данных

Как исходя из двумерной совокупности данных найти наилучший вариант линии, которая предсказывала бы Y по X? Один из подходов заключается в том, чтобы найти линию, характеризующуюся в некотором смысле наименьшей ошибкой предсказания. Удобнее всего использовать для этого **линию наименьших квадратов**, которая характеризуется наименьшей суммой квадратов ошибок предсказания (отклонений по вертикали реальных значений от линии), в сравнении с любой другой прямой линией, которую можно было бы также начертить. Для объемов продаж из табл. 11.1.1 ошибки прогнозирования, сумму квадратов которых необходимо минимизировать, показаны на рис. 11.2.5 (для линии наименьших квадратов) и на рис. 11.2.6 (для линии, выбранной не лучшим образом).

Линию наименьших квадратов построить нетрудно. Компьютеры и многие калькуляторы позволяют автоматически вычислять методом наименьших квадратов наклон, b, и сдвиг, a. Наклон иногда называют **коэффициентом регрессии** Y по X, а сдвиг — **постоянным членом**, или **константой** регрессии. Наклон, b, вычисляется как корреляция, r, умноженная на отношение стандартных откло-

[16] Линию можно определить, задав наклон и значение Y для $\bar{X}$. При таком подходе оба числа, определяющие линию, всегда имеют содержательный смысл. Однако в настоящее время так поступают редко.

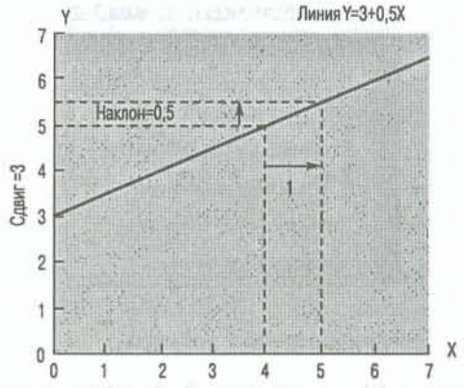


Рис. 11.2.2. Прямая линия, $Y = 3 + 0{,}5X$ начинается в точке сдвига ($a = 3$) при $X = 0$ и поднимается на 0,5 (одно значение наклона, $b = 0{,}5$) при каждом сдвиге на одну единицу вправо

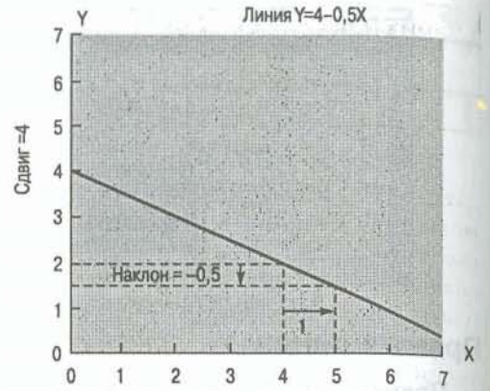


Рис. 11.2.3. Линия с отрицательным наклоном. Прямая линия, $Y = 4 - 0{,}5X$, начинается в точке сдвига ($a = 4$) при $X = 0$ и снижается на 0,5 (так как величина наклона отрицательна, $b = -0{,}5$) при каждом сдвиге на одну единицу вправо

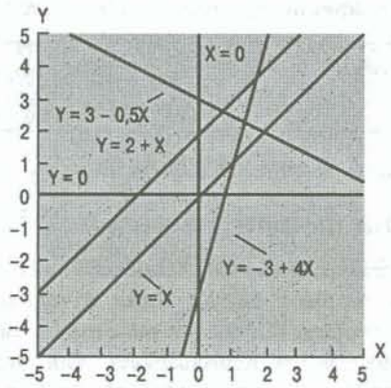


Рис. 11.2.4. Набор прямых линий и их уравнения, показывающие наклон и сдвиг. Только лишь вертикальную линию невозможно описать уравнением вида $Y = a + bX$

нений, SY/SX (выраженное в единицах Y на единицу X). Сдвиг, a, определяется таким образом, чтобы линия проходила через наиболее подходящую точку, а именно ($\bar{X}, \bar{Y}$). Соответствующие формулы имеют следующий вид.

Вычисление наклона и сдвига методом наименьших квадратов

$$\text{Наклон} = b = r\frac{S_Y}{S_X}.$$

$$\text{Сдвиг} = a = \bar{Y} - b\bar{X} = \bar{Y} - r\frac{S_Y}{S_X}\bar{X}.$$

Линия наименьших квадратов

$$(\text{Прогнозируемое значение } Y) = a + bX = (\bar{Y} - r\frac{S_Y}{S_X}\bar{X}) + r\frac{S_Y}{S_X}X.$$

Не рассчитывайте, что все точки попадут на линию. Эту линию можно считать обобщенной характеристикой взаимосвязи между переменными. Данные можно представить себе как линию с добавлением некоторой случайности. **Прогнозируемое значение** Y при заданном значении X равно высоте линии при этом значении X; такое значение Y можно вычислить исходя из уравнения линии наименьших квадратов. Прогнозируемое значение Y можно найти либо для некоторой точки имеющихся данных, либо для нового значения X. Для каждой точки имеющихся данных можно определить **остаток**, который указывает, насколько эта точка оказывается выше (или ниже — если значение остатка меньше нуля) линии. Остатки позволяют вносить определенные коррективы, сравнивая фактические значения Y с теми значениями, которые можно ожидать для соответствующих значений X. Формула вычисления остатка для точки данных (X,Y) имеет следующий вид:

остаток = (фактическое значение Y) – (прогнозируемое значение Y) = $Y - (a + bX)$.



Рис. 11.2.5. Линия наименьших квадратов характеризуется наименьшей из всех возможных линий суммой квадратов ошибок прогнозирования. Ошибки прогнозирования измеряются по вертикали



Рис. 11.2.6. Необдуманный выбор линии будет характеризоваться значительными ошибками прогнозирования и не будет соответствовать линии наименьших квадратов

Таблица 11.2.2. Недельный объем производства

Количество выпущенных изделий, X	Затраты, Y
22	3470
30	3783
26	3856
31	3910
36	4489

Окончание табл. 11.2.2

	Количество выпущенных изделий, X	Затраты, Y
	30	3876
	22	3221
	45	4579
	38	4325
	30	3589
	38	3999
	41	4158
	27	3666
	28	3885
	31	3574
	37	4495
	32	3814
	41	4430
Среднее значение	$\bar{X}$ = 32,50	$\bar{Y}$ = $3951,06
Стандартное отклонение	S_x = 6,5552	S_y = $389,6131
Корреляция	r = 0,869193	

Пример. *Фиксированные и переменные затраты*

Вернемся к данным о производстве (см. один из предыдущих примеров), но не будем рассматривать резко отклоняющееся значение. Эти данные — с указанием *X* и *Y*, а также соответствующих статистических характеристик — приведены в табл. 11.2.2. Естественно, что *X* обозначает количество произведенных изделий, а *Y* — затраты, поскольку у менеджеров часто возникает потребность прогнозировать затраты, основываясь на текущих планах выпуска продукции.

Наклон представляет *переменные затраты* (себестоимость производства еще одного изделия); его можно вычислить из имеющихся статистических характеристик следующим образом:

$$\text{переменные затраты} = b = rSY/SX = (0{,}869193)(389{,}6131)/6{,}5552 = \$51{,}66.$$

Другой член уравнения, сдвиг, определяет *фиксированные затраты*. Речь идет о таких базовых затратах, как, например, арендная плата, которая включается в расходы даже в том случае, если продукция вообще не выпускается. Сдвиг вычисляется следующим образом:[17]

$$\text{фиксированные затраты} = a = \bar{Y} - b\bar{X} = 3951{,}06 - (51{,}66)(32{,}5) = \$2272\,.$$

Выражение для линии наименьших квадратов можно представить в следующем виде:

прогнозируемые затраты =
=фиксированные затраты + (переменные затраты)(количество произведенных изделий) =
= $2272 + $51,66 (количество произведенных изделий).

[17] Чтобы иметь возможность интерпретировать вычисленное значение сдвига как фиксированные затраты, необходимо также предположить, что линейная связь присутствует даже за пределами диапазона изменений имеющихся данных, поскольку, для того, чтобы линия пересекла ось *Y* (при *X* = 0), нам необходимо продолжить построенную линию (другими словами, экстраполировать ее за пределы имеющихся у нас данных).

Данные и соответствующая линия наименьших квадратов изображены на рис. 11.2.7.

Эту оценку связи можно использовать при составлении бюджета. Если вы предполагаете, что на следующей неделе понадобится выпустить 36 изделий, то можно прогнозировать соответствующие затраты, воспользовавшись связью в данных за предыдущий период, которую отражает линия наименьших квадратов. Прогноз может быть следующим:

прогнозируемые затраты на производство 36 изделий = *a* + (*b*)(36) = $2272 + ($51,66)(36) = $4132.



Рис. 11.2.7. Линия наименьших квадратов характеризует данные о производственных затратах, определяя фиксированные затраты (сдвиг, a = $2272) и переменные затраты на одно изделие (наклон, b = $51,66 на одно произведенное изделие)

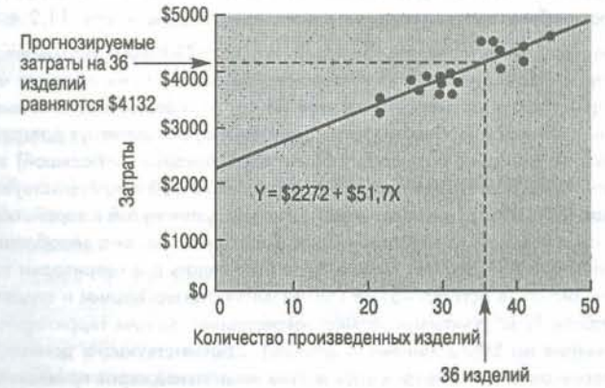


Рис. 11.2.8. Линию наименьших квадратов можно использовать для прогнозирования (или предсказания) ожидаемого значения Y, если задано новое значение X. В данном случае вы планируете выпустить на следующей неделе 36 изделий. Линия наименьших квадратов предполагает, что затраты в этом случае могут составить $4 132. Конечно, реальные затраты будут случайным образом отличаться от прогнозируемых, как, впрочем, и другие точки на диаграмме

Ваши прогнозируемые затраты равны высоте линии (по отношению к оси *X*) в точке, соответствующей выпуску 36 изделий, как показано на рис. 11.2.8. Естественно, трудно рассчитывать на то, что реальные затраты составят именно $4 132. В то же время вы можете ожидать затраты, размер которых будет лишь случайным образом отличаться от вашего наилучшего предположения $4132.

Пример. Территория и продажи

Ваши менеджеры по сбыту — вообще говоря, неравноценные работники. Разумеется, одни из них работают усерднее, чем другие, и производственные показатели (объем продаж) у них оказываются лучше, чем у других. Однако ситуация сложнее, чем может показаться на первый взгляд. За каждым из менеджеров закреплена определенная территория. Одни территории предоставляют более широкие возможности для ведения бизнеса, чем другие. Пытаясь выяснить, кто из менеджеров работает лучше, а кто — хуже, помимо анализа, сколько товара каждому из менеджеров удалось продать (что, конечно, очень важно), вы решаете сделать *поправку на размер территории*. Может оказаться, что некоторые из тех, кого вы считали хорошими работниками, достигли лучших, чем у других, показателей лишь за счет того, что за ними закреплены большие территории. Кроме того, вы можете открыть новые таланты тех менеджеров, уровень продаж которых выше среднего, но общий объем не очень высок по причине небольшой территории. Все эти поправки вам поможет внести регрессионный анализ. Соответствующая совокупность данных представлена в табл. 11.2.3.

Линия наименьших квадратов определяется следующим выражением:

ожидаемый объем продаж = $1 371 744 + $0,23675045 (территория).

Подставляя в это уравнение размер территории, закрепленной за каждым из менеджеров по сбыту, можно определить ожидаемый (в зависимости от величины территории) объем продаж. Например, у Ансона ожидаемый объем продаж составляет $1 371 744 + ($0,23675045)×(4 956 512) = $2 545 000 (результат округлен до ближайшей тысячи). Фактический объем продаж у Ансона (примерно $2 687 000) оказывается на $142 000 больше, чем ожидаемый. Таким образом, значение остатка для Ансона ($142 000) свидетельствует о его успешной деятельности. Ожидаемые объемы продаж и значения остатков можно вычислить для каждого из менеджера по сбыту; соответствующие данные представлены в табл. 11.2.4.

Остатки представляют особый интерес. Наибольший из них, $791 000, указывает, что Бонни удалось обеспечить своей фирме примерно на $0,79 миллиона больший объем продаж, чем можно было бы ожидать для территории такого размера. Несмотря на то что абсолютное значение объема продаж, обеспеченного Бонни, оказалось не самым высоким по фирме, ею достигнут достаточно впечатляющий результат, если принять во внимание размер (вообще говоря, довольно небольшой) закрепленной за ней территории. Еще один также достаточно большой остаток $538 000 свидетельствует о том, что впечатляющий объем продаж ($5 149 127 — абсолютный рекорд), достигнутый Кларой, объясняется не только большим размером закрепленной за ней территории. Действительно, она заработала для своей фирмы примерно на $0,5 миллиона больше, чем можно было бы ожидать для территории такого размера. В то же время наименьшее значение остатка –$729 000 является отрицательным и свидетельствует о серьезных недостатках в работе Рода (учитывая размер закрепленной за ним территории, он должен был бы принести фирме примерно на $0,73 миллиона больше). Соответствующие данные, линия наименьших квадратов и комментарии относительно трех упомянутых нами менеджеров приведены на рис. 11.2.9.

Не пытайтесь интерпретировать эти результаты слишком буквально. Несмотря на то что эти три особых случая действительно способны выявить двух "звезд" и одного неудачника, полученные результаты можно объяснить и по-другому. Возможно, проблемы Рода возникли потому, что его территория является одним из депрессивных регионов страны. В таком случае его относительно низкие результаты вовсе не объясняются слабыми профессиональными качествами или недостаточным усердием в работе. Возможно, требуется проведение более тщательного и сложного регрессионного анализа, который учитывал бы и другие, не менее важные факторы.

Таблица 11.2.3. Территория и производительность менеджеров по сбыту

	Территория (численность населения)	Объем продаж (за прошлый год), дол.		Территория (численность населения)	Объем продаж (за прошлый год), дол.
Ансон	4 956 512	2 687 224	Клара	13 683 663	5 149 127
Эшли	8 256 603	3 543 166	Бриттани	3 580 058	2 024 809
Джонатан	9 095 310	3 320 214	Ян	2 775 820	1 711 720
Род	12 250 809	3 542 722	Бонни	4 637 015	3 260 464
Николас	4 735 498	2 251 482			

Таблица 11.2.4. Территория, фактическая и ожидаемая производительность, а также остатки

	Территория (численность населения)	Фактический объем продаж, дол.	Ожидаемый объем продаж (округленный), дол.	Остаток (округленный), дол.
Ансон	4 956 512	2 687 224	2 545 000	142 000
Эшли	8 256 603	3 543 166	3 326 000	217 000
Джонатан	9 095 310	3 320 214	3 525 000	–205 000
Род	12 250 809	3 542 722	4 272 000	–729 000
Николас	4 735 498	2 251 482	2 493 000	–241 000
Клара	13 683 663	5 149 127	4 611 000	538 000
Бриттани	3 580 058	2 024 809	2 219 000	–195 000
Ян	2 775 820	1 711 720	2 029 000	–317 000
Бонни	4 637 015	3 260 464	2 470 000	791 000

Насколько полезна построенная линия

Вы, наверное, уже обратили внимание, что линия наименьших квадратов не является идеальным описанием данных. Она, несомненно, является полезной характеристикой основной тенденции, но все же не учитывает случайные отклонения данных от линии. В связи с этим возникает следующий вопрос: "Насколько полезна линия регрессии?" Ответ на этот вопрос основывается на двух важных показателях: *стандартной ошибки оценки* (абсолютная мера величины ошибок прогнозирования) и $R2$ (относительная мера того, как много удалось объяснить).

Стандартная ошибка оценки: насколько велики ошибки предсказания

Стандартная ошибка оценки, которую мы будем обозначать S_e (в компьютерных распечатках часто встречается обозначение S), является приближенным показателем величины ошибок предсказания (остатков) для имеющихся данных и измеряется в тех же единицах, что и Y. Насколько хорошо вы можете предска-



Рис. 11.2.9. Рассматривая положение каждой точки данных относительно линии регрессии, можно оценить производительность каждого менеджера с учетом поправки на некоторый другой фактор. В данном случае точки, расположенные над линией (имеющие положительный остаток), представляют менеджеров, обеспечивших более высокие объемы продаж, чем можно было бы ожидать, учитывая размер закрепленных за ними территорий. Точки, расположенные под линией, представляют объемы продаж ниже ожидаемых

зывать Y? Ответ: с точностью плюс-минус несколько S_e.[18] Поскольку, как правило, требуется, чтобы прогноз был как можно более точным, значение S_e должно быть как можно *меньшим*. S_e можно интерпретировать как стандартное отклонение в том смысле, что если ошибки предсказания имеют нормальное распределение, то можно ожидать, что примерно 2/3 точек данных будут находиться на расстоянии не более S_e выше или ниже линии регрессии. Кроме того, около 95% значений данных должны находиться на расстоянии не более чем $2S_e$ от линии регрессии и т.д. Рис. 11.2.10 иллюстрирует это положение на примере данных о производственных затратах.

Стандартную ошибку оценки можно вычислить с помощью следующих формул.

Стандартная ошибка оценки

$$S_e = S_Y\sqrt{(1-r^2)\frac{n-1}{n-2}} \quad \text{(для вычисления)}$$

$$= \sqrt{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n}[Y_i-(a+bX_i)]^2} \quad \text{(для интерпретации)}.$$

Первая формула показывает, как вычислять S_e путем уменьшения S_Y с учетом корреляции и размера выборки. Действительно, S_e, как правило, меньше S_Y, поскольку линия $a + bX$ характеризует соответствующую взаимосвязь и, следовательно, оказывается ближе к значениям Y, чем более простая статистическая характеристика, $\overline{Y}$. Вторая формула показывает способ интерпретации S_e как оценки стандартного отклонения остатков: возведенные в квадрат ошибки прогнозирования усредняются путем деления на $n - 2$ (соответствующее количество степеней свободы, поскольку выполнена оценка двух чисел, a и b), а корень квадратный позволяет получить результат в тех же единицах, в которых измерена переменная Y.

[18] Более строгий, точный ответ на этот вопрос мы дадим в одном из следующих разделов, когда речь пойдет о прогнозировании нового значения Y для заданного значения X.

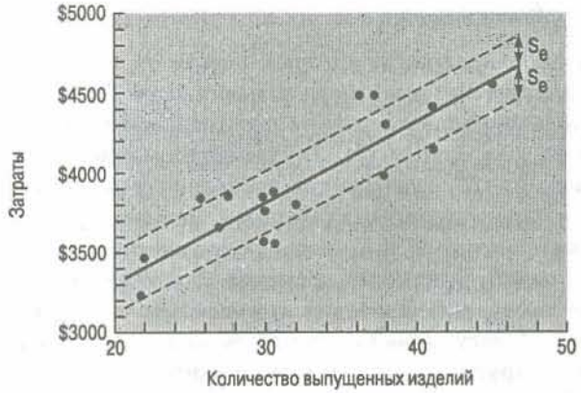


Рис. 11.2.10. Стандартное отклонение оценки, S_e, показывает приблизительно, ошибку какой величины вы допускаете, когда вместо фактического значения Y используете прогнозируемое (на линии наименьших квадратов) значение Y. Можно ожидать, что в случае обычной линейной связи (такой, как показана на этом рисунке) примерно 2/3 точки данных будет находиться выше или ниже линии наименьших квадратов на расстоянии не более S_e

В задаче о производственных затратах мы имели коэффициент корреляции $r = 0{,}869193$, изменчивость в отдельных значениях затрат $S_Y = \$389{,}6131$ и размер выборки $n = 18$. В таком случае стандартная ошибка оценки равна

$$S_e = S_Y\sqrt{(1-r^2)\frac{n-1}{n-2}} = 389{,}6131\sqrt{(1-0{,}869193^2)\frac{18-1}{18-2}} =$$

$$= 389{,}6131\sqrt{(0{,}0244503)\frac{17}{16}} = 389{,}6131\sqrt{0{,}259785} = \$198{,}58\,.$$

Этот результат свидетельствует о том, что в обычные недели фактические затраты отличались от прогнозируемых (методом наименьших квадратов) затрат примерно на \$198,58. Несмотря на то что линия наименьших квадратов (линия прогноза) максимально учитывает взаимосвязь между затратами и объемом произведенной продукции, прогнозы, которые можно делать на основе этой линии, далеки от совершенства.

R^2: как много объяснено

R^2 (произносится "r-квадрат"), который называют также **коэффициентом детерминации**, показывает, в какой мере изменчивость Y объясняется поведением X.[19] Этот показатель вычисляется путем простого возведения в квадрат коэффициента корреляции, r (т.е. $R^2 = r2$). Таким образом, доля вариации Y, определяемая выражением $1 - R^2$, оказывается необъясненной. Обычно большие значения R^2 считаются более предпочтительными, поскольку указывают на более сильную взаимосвязь между X и Y, которую можно использовать для прогнозирования и других целей. Однако на практике малые значения R^2 вовсе необязательно указывают на то, что X нельзя использовать для объяснения поведения Y; малые значения R^2 могут просто указывать на то, что поведение Y объясняется не только X, но и другими важными факторами.

Например, коэффициент корреляции совокупности данных, относящихся к производственным затратам, равняется 0,869193. Следовательно, значение R^2 равно

$$R^2 = 0{,}869193^2 = 0{,}755, \text{ или } 75{,}5\%.$$

Это значение R^2 говорит о том, что 75,5% вариации (изменчивости) недельных затрат объясняется количеством изделий, выпущенных за неделю. Остальная часть (24,5%) вариации общих затрат объясняется другими причинами.

Можно представить себе это таким образом. Каждую неделю наблюдается определенная вариация (изменчивость) величины производственных затрат (которая характеризуется показателем S_Y). Часть этой вариации объясняется тем, что в какие-то недели уровень производства оказывается выше (что и приводит к более высоким затратам), а в какие-то недели — ниже. Таким образом, кличество произведенных изделий "объясняет" определенную часть вариации недельного уровня затрат. Однако это не позволяет понять *всю* вариацию. Существуют и другие факторы (например, неожиданные поломки оборудования, сверхурочные работы, определенные ошибки и т.п.), которые также сказываются на вариации уровня затрат. Такое значение R^2 свидетельствует о том, что 75,5% вариации недельных затрат можно отнести на счет объема производства, а оставшиеся 24,5% вариации все еще не объяснены.

Доверительные интервалы и проверка гипотез для регрессии

До сих пор мы занимались обобщением *данных*: оценивали силу взаимосвязи с помощью коэффициента корреляции, взаимосвязь с помощью линии наименьших квадратов, соответствие линии и данных с помощью стандартной ошибки оценки и R^2. Сейчас настало время сделать следующий шаг и перейти от вычисления характеристик данных выборки к статистическим выводам относительно более крупной генеральной совокупности, которая нас, собственно, и интересует. Но что необходимо рассматривать в случае регрессии как генеральную совокупность? Ответ на этот вопрос дает *линейная модель*.

[19] Более точно, R^2 является той частью *дисперсии* Y, которая объясняется влиянием X. По определенным техническим причинам (поскольку *квадрат* полной ошибки можно представить в виде *квадратов* двух компонентов: объясненной части и необъясненной) в статистике традиционно используют дисперсию (квадрат стандартного отклонения).

Предположение о линейности определяет генеральную совокупность

Чтобы статистический вывод был обоснованным, анализируемые данные должны представлять собой случайную выборку из интересующей нас генеральной совокупности. Как всегда, это гарантирует, что данные точным и предсказуемым образом представляют интересующую нас генеральную совокупность. Кроме того, нам нужно сделать определенное техническое допущение, которое позволит использовать t-таблицу, в основе которой лежит нормальное распределение. С этой целью мы будем предполагать, что данные для обеих переменных извлечены независимо и соответствуют **линейной модели**, которая утверждает, что наблюдаемое значение Y определяется в генеральной совокупности характеризующейся прямой линией связью плюс случайная, имеющая нормальное распределение, ошибка.

Линейность генеральной совокупности (линейная модель)

$$Y = (\alpha + \beta X) + \varepsilon =$$

$$= (\text{связь в генеральной совокупности}) + \text{случайность;}$$

где ε имеет нормальное распределение со средним значением 0 и постоянным стандартным отклонением σ.

Эти допущения обеспечивают дополнительные гарантии того, что выбранная совокупность данных будет состоять из независимых наблюдений, характеризующихся линейной связью с одинаковой вариацией и приблизительно нормально распределенной случайностью.

Связь в генеральной совокупности задается двумя параметрами: сдвигом (константа уравнения регрессии) в генеральной совокупности α и наклоном (коэффициент регрессии) в генеральной совокупности β. Еще один параметр генеральной совокупности, σ, указывает величину неопределенности в этой ситуации. Если бы ваши данные представляли, например, данные переписи всего населения, тогда соответствующая линия наименьших квадратов представляла бы связь в генеральной совокупности. Как правило, в качестве *оценки* α используют вычисленный с помощью метода наименьших квадратов сдвиг a; в качестве *оценки* β — вычисленный методом наименьших квадратов наклон b; в качестве *оценки* σ — стандартную ошибку оценки, S_e. Разумеется, со всеми этими оценками связаны определенные ошибки, поскольку a, b и S_e вычисляются на небольших выборках, а не на всей генеральной совокупности. В табл. 11.2.5 представлена сводка этих параметров генеральной совокупности и выборочных статистик.

Линейность является базовым допущением для статистических выводов в регрессионном и корреляционном анализе. Построение доверительных интервалов и проверка статистических гипотез для коэффициента регрессии предполагают, что линейность справедлива для генеральной совокупности. В частности, доверительные интервалы и проверки гипотез будут необоснованны, если соответствующая взаимосвязь окажется нелинейной или будет характеризоваться неодинаковой вариацией. Вам необходимо учитывать эти особенности: если линейная модель не соответствует вашим данным, то выводы, сделанные на основе регрессионного анализа, могут оказаться неверными.

Стандартные ошибки для наклона и сдвига

Когда речь идет о параметрах генеральной совокупности и о выборочных оценках, естественно предположить, что где-то скрыты соответствующие стандартные ошибки. Зная эти стандартные ошибки и количество степеней свободы, можно использовать уже известные из глав 9 и 10 методы для построения доверительных интервалов и проверки статистических гипотез.

Стандартная ошибка коэффициента регрессии, S_b, указывает приблизительную величину вызванного случайностью выборки отклонения оценки наклона, b (коэффициент регрессии, вычисленный на основе выборки), от наклона в генеральной совокупности, β. Обратите внимание, что S_b является выборочной статистикой. Формула для S_b выглядит следующим образом.

Стандартная ошибка коэффициента регрессии

$$S_b = \frac{S_e}{S_X\sqrt{n-1}}$$; число степеней свободы равно $n - 2$.

Эта формула свидетельствует о том, что неопределенность b пропорциональна базовой неопределенности (Se) в данной ситуации, но (1) S_b будет меньше, когда значение SX оказывается большим (поскольку линия будет определена лучше, если диапазон значений X будет больше), и (2) S_b будет меньше, когда размер выборки n будет больше (просто потому, что в этом случае у нас оказывается больше информации). Довольно часто используют такие термы, как корень квадратный из n в знаменателе формулы стандартной ошибки, которые отражают влияние дополнительной информации.

Число степеней свободы для этой стандартной ошибки равняется $n - 2$, поскольку при построении линии регрессии оцениваются два значения, a и b.

В нашем примере с производственными затратами (без выброса!) для выборочных данных имеем коэффициент корреляции $r = 0{,}869193$, размер выборки $n = 18$ и наклон (переменные затраты) $b = 51{,}66$. Мы имеем дело с идеализированной генеральной совокупностью, состоящей из всех тех недель, которые обстоятельствами и условиями работы ничем не отличаются от той недели, которую мы наблюдали. Тогда можно считать, что коэффициент регрессии в генеральной совокупности, β, равен тому коэффициенту регрессии, который вы могли бы вычислить, если бы у вас в распоряжении было намного больше данных. Стандартная ошибка b равна:

$$S_b = \frac{S_e}{S_X\sqrt{n-1}} = \frac{198{,}58}{6{,}5552\sqrt{18-1}} = \frac{198{,}58}{27{,}0278} = 7{,}35\,.$$

Таблица 11.2.5. Параметры генеральной совокупности и выборочные статистики

	Генеральная совокупность (параметры: фиксированные и неизвестные)	Выборка (оценки: случайные и известные)
Сдвиг	α	a
Наклон	β	b
Линия регрессии	$Y = \alpha + \beta X$	Y = a + bX
Неопределенность	σ	Se

Исходя из этих же данных была вычислена оценка сдвига a. Следовательно, этот параметр также характеризуется стандартной ошибкой, указывающей на неопределенность его оценки. **Стандартная ошибка сдвига**, Sa, указывает приблизительно, насколько далеко оценка a отстоит от α, истинной величины сдвига в генеральной совокупности. Эта стандартная ошибка, формула для вычисления которой приведена ниже, также имеет $n - 2$ степеней свободы и представляет собой выборочную статистику.

Стандартная ошибка сдвига

$$S_a = S_e\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_X^2(n-1)}}$$; число степеней свободы равно $n - 2$.

Эта формула указывает на то, что неопределенность a пропорциональна базовой неопределенности (Se), что неопределенность a уменьшается при увеличении размера выборки n и увеличивается, когда абсолютное значение $\bar{X}$ увеличивается в сравнении с SX (поскольку данные по X будут далеко отстоять от 0 — точки, в которой определяется сдвиг), и что существует "базовый" член $1/n$, поскольку a было бы средним значением Y при $X = 0$.

В нашем примере с производственными затратами сдвиг $a = \$2272$ служит оценкой фиксированных затрат. Стандартная ошибка этой оценки равна:

$$S_a = S_e\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_X^2(n-1)}} =$$

$$= 198{,}58\sqrt{\frac{1}{18} + \frac{32{,}50^2}{6{,}5552^2(18-1)}} =$$

$$= 198{,}58\sqrt{0{,}0555556 + \frac{1056{,}25}{730{,}50}} =$$

$$= 198{,}58\sqrt{1{,}5015} = 243{,}33\,.$$

Доверительные интервалы для коэффициентов регрессии

Этот материал должен быть уже вам знаком. Вы берете оценку (например, b), ее собственную стандартную ошибку (например, S_b) и t-значение из t-таблицы (для $n - 2$ степеней свободы). Двусторонний доверительный интервал находится между $b - tS_b$ и $b + tS_b$. Односторонний доверительный интервал утверждает либо то, что наклон в генеральной совокупности, β, будет не меньше, чем $b - tS_b$, или что наклон в генеральной совокупности, β, либо то, что он будет не больше, чем $b + tS_b$ (с использованием, разумеется, соответствующих односторонних t-значений). Чтобы восстановить в памяти основные понятия, касающиеся доверительных интервалов, вам, возможно, придется еще раз обратиться к резюме главы 9; единственное отличие в данном случае заключается в том, что вы оцениваете не среднее значение в генеральной совокупности, а скорее *взаимосвязь* в генеральной совокупности.

Таким же образом вывод о сдвиге α основывается на оценке a и ее стандартной ошибке Sa.

Доверительные интервалы

Для наклона (коэффициента регрессии) в генеральной совокупности, β:

$$\text{от } b - tS_b \text{ до } b + tS_b.$$

Для сдвига в генеральной совокупности, α:

$$\text{от } a - tS_a \text{ до } a + tS_a.$$

Пример. *Переменные затраты производства*

Для данных о производственных затратах оценка наклона $b = 51{,}66$, ее стандартная ошибка $S_b = 7{,}35$, а двустороннее значение из t-таблицы для $n - 2 = 16$ степеней свободы на доверительном уровне 95% равно 2,120. Таким образом, 95% доверительный интервал для β находится между $51{,}66 - (7{,}35)(2{,}120) = 36{,}08$ и $51{,}66 + (7{,}35)(2{,}120) = 67{,}24$. Утверждение о доверительном интервале в этом случае будет выглядеть так.

"Мы на 95% уверены в том, что в долгосрочном плане (для генеральной совокупности) размер переменных затрат будет находиться между \$36,08 и \$67,24 на каждое произведенное изделие".

Как это часто бывает, доверительный интервал лишний раз напоминает нам о том, что оценка (\$51,66) является весьма приблизительной. Рассматривая свои данные как случайную выборку из генеральной совокупности объемов продукции и затрат, которые имели место ранее при сходных обстоятельствах, мы приходим к заключению, что, имея данные лишь за 18 прошлых недель, мы получаем существенную неопределенность в размере переменных затрат.

Односторонний доверительный интервал дает нам достаточно обоснованную верхнюю границу, которую можно использовать при составлении бюджета. Это отражает факт вашего незнания действительного размера переменных затрат: в вашем распоряжении имеется лишь их *оценка*. В этом примере одностороннее t-значение из таблицы равно 1,746, поэтому верхняя граница будет равна $51{,}66 + (7{,}35)(1{,}746) = 64{,}49$. Утверждение об одностороннем доверительном интервале будет иметь следующий вид.

"Мы на 95% уверены в том, что в долгосрочном плане (для генеральной совокупности) размер переменных затрат не превысит \$64,49 на каждое произведенное изделие".

Обратите внимание, что эта граница (\$64,49) оказывается меньше, чем верхняя граница двустороннего интервала (\$67,24), поскольку вас интересует только эта (верхняя) сторона интервала. Таким образом, поскольку нас совершенно не интересует нижняя сторона интервала, мы получили верхнюю границу, которая ближе к значению оценки \$51,66.

Проверка того, является связь реальной или случайной

Эта глава посвящена взаимосвязи между X и Y. Корреляция характеризует силу этой взаимосвязи, а уравнение регрессии использует эту взаимосвязь для предсказания поведения Y по X. Однако, как это нередко случается в статистике, можно вычислять характеристики связи даже тогда, когда в действительности ее нет. Задача проверки гипотез в том и заключается, чтобы выяснить, является ли взаимосвязь, которая, *как вам кажется*, присутствует в данных, чистой случайностью или отражает реальную и значимую связь между X и Y.

Нулевая гипотеза утверждает, что между X и Y *никакой взаимосвязи нет* и что выявленная нами взаимосвязь в данных — не что иное, как продукт случайного сочетания определенных пар значений X и Y. Единственный вариант, когда в рамках линейной модели $Y = \alpha + \beta X + \varepsilon$ Y *не* зависит реально от X, имеет место лишь тогда, когда $\beta = 0$, т.е. когда X исчезает и линейная модель сводится к $Y = \alpha + \varepsilon$. Еще один способ сказать, что взаимосвязь между X и Y отсутствует, заключается в том, чтобы сказать, что X и Y *независимы* друг от друга.

Альтернативная (исследовательская) гипотеза утверждает, что между X и Y действительно *существует взаимосвязь*, которая не является случайностью. Это возможно тогда, когда $\beta \neq 0$, т.е. в линейной модели для Y сохраняется составляющая, зависящая от X. Математическая запись этих гипотез имеет следующий вид.

Гипотезы для проверки значимости взаимосвязи

$$H_0: \beta = 0;$$

$$H_1: \beta \neq 0.$$

Сама по себе эта проверка выполняется обычным способом — ничего нового для вас нет и в этом случае.[20] Можно использовать метод доверительного интервала, чтобы выяснить, попадает ли в доверительный интервал заданное значение 0, и, если не попадает, принять решение о значимости взаимосвязи (принять $H1$). Или можно вычислить t-статистику b/S_b, сравнить ее с t-значением из таблицы и принять решение о значимости взаимосвязи (принять $H1$), если абсолютное значение t-статистики окажется больше.

Вернемся к примеру о переменных производственных затратах. Доверительный интервал в этом случае находится между \$36,08 и \$67,24. Поскольку заданное значение 0 в доверительный интервал не попадает, можно сделать вывод о том, что мы имеем дело со значимыми переменными затратами. То есть, исходя из имеющихся у нас данных, можно сказать, что между количеством произведенных в течение недели изделий и затратами действительно существует взаимосвязь. Столь очевидную зависимость (чем больше количество произведенной продукции, тем, как правило, выше затраты) невозможно объяснить одной лишь случайностью.

Разумеется, подход, основанный на t-статистике, дает тот же ответ. t-статистика в нашем случае определяется как $t = b/S_b = 51{,}66/7{,}35 = 7{,}03$. Поскольку абсолютное значение t-статистики (7,03) оказывается больше, чем значение из t-таблицы (2,120) с $n - 2 = 16$ степенями свободы при проверке на уровне 5%, то можно сделать вывод о том, что коэффициент регрессии (51,66) действительно значимо отличается от 0.

Другие методы проверки значимости взаимосвязи

Существуют и другие методы проверки значимости взаимосвязи. Несмотря на то что на первый взгляд может показаться, что они существенно отличаются от описанного выше, ответ, полученный с их помощью, в любом случае будет таким же, как в описанных выше методах, основанных на коэффициенте регрессии. Эти альтернативные методы основаны на других статистических характеристиках: например, на коэффициенте корреляции, r, а не на коэффициенте наклона, b. Но поскольку основной вопрос остается тем же (есть взаимосвязь или нет?), ответы на него в любом случае также будут одними и теми же. Это можно доказать математически.

[20] Если вам требуется освежить в памяти основы проверки гипотез, обратитесь к резюме главы 10.

Существуют два способа проверить значимость исходя из коэффициента корреляции. Зная коэффициент корреляции, можно обратиться к специальной таблице или преобразовать коэффициент корреляции и найти t-статистику $t = r\sqrt{(n-2)/(1-r^2)}$, которую затем сравнить со значением из t-таблицы с $n-2$ степенями свободы. В конечном счете эти методы позволяют получить тот же ответ, что и проверка с помощью коэффициента наклона. Фактически t-статистика, определенная с помощью коэффициента корреляции, имеет то же значение, что и t-статистика, определенная с помощью коэффициента наклона ($t = b/S_b$).

Это означает следующее: вы можете прийти к выводу о наличии значимой (или, наоборот, незначимой) корреляции, основываясь на проверке значимости коэффициента регрессии, b. Фактически мы делаем вывод о наличии значимой положительной корреляции, если соответствующая взаимосвязь является значимой и $b > 0$. Или, если эта взаимосвязь является значимой, а $b < 0$, мы делаем вывод о наличии значимой отрицательной корреляции.

Есть специальная проверка значимости, называемая F-тестом, которая позволяет оценить суммарную значимость регрессионной связи. Мы рассмотрим эту проверку позже, в главе о множественной регрессии. Несмотря на то что эта проверка, на первый взгляд, также существенно отличается от описанных выше, в конечном счете она сводится к тому же, что и проверка на основе коэффициента наклона, когда есть только X и Y и никакие другие переменные не рассматриваются.

Результаты компьютерных вычислений для данных о производственных затратах

Многие из полученных нами результатов, касающихся данных о производственных затратах, можно получить с помощью компьютера. Первым на компьютерной распечатке выводится уравнение прогноза ("Уравнение регрессии"). Далее выводятся коэффициенты ("Коэфф"), $a = 2272{,}1$ и $b = 51{,}661$ со своими стандартными ошибками ("СтнОш"), $S_a = 243{,}3$ и $S_b = 7{,}347$, своими t-статистиками, $t_a = 9{,}34$ и $t_b = 7{,}03$, и своими p-значениями (оба коэффициента чрезвычайно высоко значимы, поскольку $p < 0{,}001$ в обоих случаях). Затем выводится стандартная ошибка оценки, $Se = 198{,}6$ и $R^2 = 0{,}755$.

Уравнение регрессии:
затраты = 2272 + 51,7 объем производства

Независимая переменная	Коэфф	СтнОш	t-коэффициент	p
Константа	2272,1	243,3	9,34	0,000
Объем производства	51,661	7,347	7,03	0,000

S = 198,6 R-sq = 75,5% R-sq (коррект.) = 74,0%.

Пример. *Рассмотрим еще раз инерцию фондовой биржи*

Ранее в этой главе суточные изменения процентов на фондовой бирже использовались в качестве примера явного отсутствия взаимосвязи между X = вчерашнее изменение и Y = сегодняшнее изменение. Попытаемся воспользоваться регрессией, чтобы оценить взаимосвязь между вчерашним и сегодняшним изменениями, а затем воспользуемся проверкой гипотез, чтобы выяснить, является ли эта взаимосвязь значимой. Соответствующая совокупность данных (с линией наименьших квадратов) показана на рис. 11.2.11.

Линия наименьших квадратов определяется следующим выражением:

сегодня = 0,0003984 + 0,111421(вчера).

Например, 30 июня 1998 г. $X = 0{,}47\% = 0{,}0047$, а $Y = -0{,}41\% = -0{,}0041$. Прогнозируемое на этот день значение Y равняется $0{,}0003984 + 0{,}111421 \times 0{,}0047 = 0{,}00092$, или 0,092%.

Следует ли доверять этому "уравнению прогноза"? Вообще говоря, задача этого уравнения — помочь вам прогнозировать нынешнее поведение фондовой биржи, основываясь на ее вчерашнем поведении (предполагая, что фондовая биржа продолжает вести себя так, будто данные, характеризующие это поведение, взяты из той же генеральной совокупности). Ключевым является коэффициент наклона $b = 0{,}1114$, который свидетельствует о том, что в среднем вчерашний подъем (или падение) лишь приблизительно на 11% продолжится и сегодня. Однако насколько точно нам удалось оценить величину этого коэффициента? Ответ на этот вопрос можно найти, обратившись к доверительному интервалу, основанному на этой оценке ($b \approx 0{,}1114$), ее стандартной ошибке ($S_b = 0{,}1522$) и значении из t-таблицы с $42 - 2 = 40$ степенями свободы (что составляет 1,960 в t-таблице для бесконечного числа степеней свободы, но мы воспользуемся более точным значением, 2,02, полученным с помощью компьютера). В этом случае можно сформулировать следующее утверждение о доверительном интервале.

"Мы на 95% уверены в том, что в генеральной совокупности значение наклона β находится в диапазоне от –0,196 до 0,419".



Рис. 11.2.11. Суточные изменения процентов на фондовой бирже (X = вчера и Y = сегодня) с 1 мая по 30 июня 1998 г. Линия наименьших квадратов почти горизонтальна (с небольшим отклонением). Поскольку это небольшое отклонение может объясняться фактором случайности, проверка гипотез позволяет сделать вывод об отсутствии значимой связи между вчерашним и сегодняшним поведением фондовой биржи

Это довольно широкий интервал; более того, он включает 0, что указывает на отсутствие взаимосвязи. Таким образом, мы приходим к выводу, что, поскольку этот интервал включает 0, отмеченный нами наклон *не является значимым*, т.е. значимой связи между вчерашним и нынешним поведением фондовой биржи нет. Можно также сказать, что нет значимого отличия значения коэффициента наклона от 0.

Подход, основанный на использовании *t*-статистики, дает, разумеется, тот же ответ. Стандартная ошибка коэффициента регрессии $S_b = 0{,}1522$, поэтому *t*-статистика равняется:

$$t = b/S_b = 0{,}1114/0{,}1522 = 0{,}732.$$

При столь малом значении *t*-статистика является незначимой (сравниваем со значением 1,960 из *t*-таблицы или с более точным значением 2,02 для 40 степеней свободы).

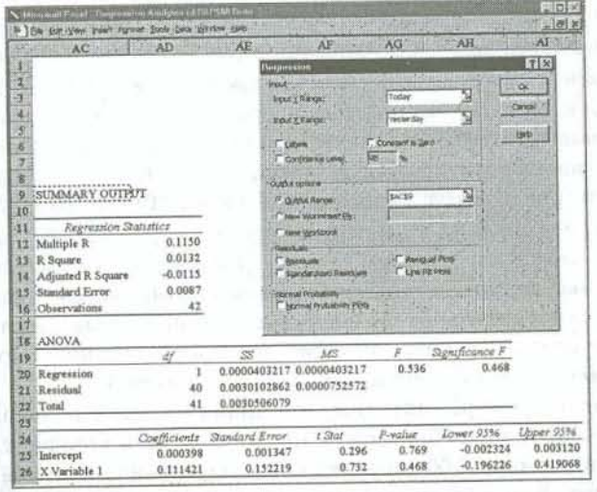
Чтобы выполнить регрессионный анализ с помощью Excel, нужно сначала присвоить имя каждому столбцу чисел, воспользовавшись командой меню Excel Insert⇨Name⇨Define (Вставка⇨Имя⇨Присвоить). Затем в меню Tools (Сервис) нужно выбрать команду Data Analysis (Анализ данных),[21] а затем пункт Regression (Регрессия). В диалоговом окне, которое появится на экране, можно указать имя входного интервала для переменной *Y* (в нашем примере — "Today") и для переменной *X* (в нашем примере — "Yesterday"). Выберите в диалоговом окне переключатель Output Range (Выходной диапазон) и укажите, в каком месте рабочего листа вы хотите поместить результаты; затем щелкните на кнопке OK. На следующем рисунке приведено диалоговое окно для этого примера и полученные результаты, среди которых можно увидеть значение R^2=0,0132 (или 1,32%), стандартную ошибку оценки S_e=0,0087, а также $b = 0{,}1114$, $S_b = 0{,}1522$, $t = 0{,}732$ и *p*-значение, равное 0,468.

Проверки других гипотез о коэффициенте регрессии

В некоторых приложениях может потребоваться проверить, насколько коэффициент регрессии отличается от некоторого *заданного значения* β_0, выполняющего роль внешнего стандарта для сравнения. Источником такого заданного значения не может служить та же совокупность данных, которая используется для регрессии. Например, вы хотите проверить, в какой мере полученные вами недавно переменные затраты (наклон в регрессии Y = затраты на X = количество произведенных изделий) отличаются от тех, которые вы использовали при составлении сметы в прошлом году (заданное значение).

Проверка значимости взаимосвязи между X и Y, речь о которой шла в предыдущем разделе, на самом деле сводится к проверке значимости отличия величины наблюдаемого наклона b от заданного значения $\beta_0 = 0$, которое выражает условие отсутствия взаимосвязи. В этом разделе мы допускаем, что β_0 может быть ненулевой величиной. Проверка выполняется обычным способом. Гипотезы и результаты имеют следующий вид.

[21] Если в меню Tools (Сервис) отсутствует пункт Data Analysis (Анализ данных), то сначала убедитесь, что вы выбрали ячейку электронной таблицы (а не график, например). Если вы все же не можете найти Data Analysis (Анализ данных), поищите пункт меню Add-Ins (Надстройки) и поставьте отметку возле Analysis ToolPak (Пакет анализа). Если это не поможет, то, видимо, необходимо переустановить Excel.



SUMMARY OUTPUT

Regression Statistics

Multiple R	0.1150
R Square	0.0132
Adjusted R Square	-0.0115
Standard Error	0.0087
Observations	42

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.0000403217	0.0000403217	0.536	0.468
Residual	40	0.0030102862	0.0000752572		
Total	41	0.0030506079			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.000398	0.001347	0.296	0.769	-0.002324	0.003120
X Variable 1	0.111421	0.152219	0.732	0.468	-0.196226	0.419068

Нулевая и альтернативная гипотезы для проверки значения коэффициента регрессии

Двусторонняя проверка:

$$H_0: \beta = \beta_0;$$
$$H_1: \beta \neq \beta_0.$$

Односторонняя проверка:

$$H_0: \beta \leq \beta_0;$$
$$H_1: \beta > \beta_0.$$

или

$$H_0: \beta \geq \beta_0;$$
$$H_1: \beta < \beta_0.$$

Результаты проверки

Если β_0 *не попадает* в доверительный интервал для коэффициента регрессии, то полученный результат является *значимым*. В случае двусторонней проверки используется двусторонний интервал и делается вывод, что b значимо отличается от β_0. Если коэффициент b больше, чем β_0, можно сделать вывод, что он значимо больше; в противном случае он значимо меньше. В случае односторонней проверки используется односторонний доверительный интервал и делается вывод, что b либо значимо больше, либо значимо меньше, чем β_0, — по обстоятельствам.

Если β_0 *попадает* в доверительный интервал для коэффициента регрессии, то полученный результат *не является значимым*. В случае двусторонней проверки используется двусторонний интервал и делается вывод, что b значимо не отличается от β_0. В случае односторонней проверки используется односторонний доверительный интервал и делается вывод, что b либо не является значимо большим, либо не является значимо меньшим, чем β_0, в зависимости от выдвинутой гипотезы.

Разумеется, можно использовать и *t*-тест. *t*-статистика определяется следующим образом:

$$t = \frac{b - \beta_0}{S_b}.$$

С помощью t-статистики можно проверить эти гипотезы о коэффициенте регрессии в генеральной совокупности β точно так же, как это было сделано в главе 10 для одно- и двусторонней проверки относительно среднего генеральной совокупности, μ.

Возвращаясь к нашему примеру о переменных производственных затратах, допустим, что ваш процесс составления сметы предполагает переменные затраты в размере \$100 на каждое произведенное изделие. Вычисленный ранее 95% доверительный интервал затрат простирается от \$36,08 до \$67,24 на каждое произведенное изделие. Поскольку заданное значение, β_0 = \$100, не попадает в этот доверительный интервал, можно сделать вывод, что вычисленная величина оценки переменных затрат, b = \$51,66, значимо отличается от предполагаемой в вашей смете. В действительности, поскольку оцениваемые затраты оказываются меньше, можно сделать вывод, что фактическая величина переменных затрат будет *значимо меньше того, что предполагалось сметой*.

Продолжая этот пример, допустим, что знания и опыт подсказывают вам, что один из ваших конкурентов борется за право получить контракт исходя из переменных затрат в размере \$60 на каждое произведенное изделие. Поскольку это заданное значение, β_0 = \$60, попадает в доверительный интервал для ваших переменных затрат, о значимой разнице говорить не приходится. Можно сделать вывод, что ваши переменные затраты значимо не отличаются от переменных затрат конкурента. Несмотря на то что ваши оцениваемые переменные затраты (\$51,66) ниже, это вполне может объясняться действием фактора случайности, а не какими-либо реальными преимуществами в затратах.

Новое наблюдение: неопределенность и доверительный интервал

Когда вы используете регрессию, чтобы сделать прогноз относительно значения нового наблюдения, желательно знать связанную с этим неопределенность. Возможно, потребуется даже сформировать соответствующий доверительный интервал, о котором известно, что он с вероятностью 95% включает следующее наблюдаемое значение.

В этой ситуации вам известно значение X_0, и вы прогнозируете значение $a + bX_0$ для Y. Есть два источника неопределенности, объединив которые можно найти стандартную ошибку прогноза. Во-первых, поскольку a и b представляют собой оценки, предсказанное значение $a + bX_0$ содержит элемент неопределенности. Во-вторых, всегда присутствует элемент случайности, ε, являющийся частью линейной модели (со стандартным отклонением, которое оценивается стандартной ошибкой S_e), и эту случайность следует учитывать, когда вы анализируете отдельное наблюдение. Результатом сочетания этих неопределенностей является стандартная ошибка Y при заданном значении X_0, обозначаемая как $S_{(Y|X_0)}$.[22] Ниже приведены соответствующая формула (и соответствующее число степеней свободы), а также результирующее определение доверительного интервала.

[22] Обратите внимание на использование оборота *при заданном значении*; эта ситуация подобна случаю *условной вероятности*, когда для уточнения вероятности вы используете дополнительную информацию. Когда вам известно значение X_0, вы располагаете дополнительной информацией, которую можно использовать для того, чтобы уменьшить неопределенность Y от (безусловного) стандартного отклонения S_Y до условной стандартной ошибки $S_{(Y|X_0)}$.

Стандартная ошибка нового наблюдения Y при заданном значении X_0

$S_{(Y|X_0)} = \sqrt{S_e^2\left(1+\frac{1}{n}\right)+S_b^2(X_0-\bar{X})^2}$; число степеней свободы равно $n-2$.

Доверительный интервал для нового наблюдения Y при заданном значении X_0

От $(a+bX_0)-tS_{(Y|X_0)}$ до $(a+bX_0)+tS_{(Y|X_0)}$.

Стандартная ошибка зависит от S_e (базовая неопределенность рассматриваемой ситуации), от S_b (неопределенность в наклоне, используемом для прогнозирования) и от расстояния между X_0 и $\bar{X}$. Стандартная ошибка нового наблюдения будет меньше, когда X_0 близко к $\bar{X}$, поскольку именно в этом случае вы знаете больше. Стандартная ошибка нового наблюдения будет большой, когда X_0 отстоит далеко от $\bar{X}$, поскольку информация, которой вы располагаете (наблюдаемые значения X), недостаточно точна в сравнении с той информацией, которая вам требуется (X_0). Эти особенности представлены на рис. 11.2.12 для совокупности данных, отражающих производственные затраты.

Возвращаясь к нашему примеру с производственными затратами, допустим, что вы запланировали выпустить на следующей неделе X_0 = 39 изделий. Воспользовавшись уравнением прогноза, вы оцениваете затраты следующим образом: $a + bX_0 = 2272 + (51,66)(39) = \$4287$. Неопределенность в этой оценке затрат на следующую неделю вычисляется по формуле

$$S_{(Y|X_0)} = \sqrt{S_e^2\left(1+\frac{1}{n}\right)+S_b^2(X_0-\bar{X})^2} =$$

$$= \sqrt{198,58^2\left(1+\frac{1}{18}\right)+7,35^2(39-32,50)^2} =$$

$$= \sqrt{(39434)(1,055556)+(54,0225)(42,25)} = \sqrt{43907} = \$209,54.$$

95% доверительный интервал на основе t-значения (2,120) из t-таблицы при $n - 2 = 16$ степенях свободы простирается от \$4287 – (2,120)(209,54) до \$4287 + (2,120)(209,54). Следовательно:

"Мы на 95% уверены в том, что на следующей неделе производственные затраты (прогнозируемые на уровне \$4287 для выпуска 39 изделий) окажутся в диапазоне от \$3843 до \$4731".

Этот доверительный интервал учитывает все статистические источники ошибок: незначительный размер выборки, оценку линии наименьших квадратов для прогнозирования и оценочную величину дополнительной неопределенности нового наблюдения. Если линейная модель является адекватным описанием вашей структуры затрат, тогда доверительный интервал будет правильным. Этот статистический метод, однако, не может учитывать другие источники ошибок, такие как возможный пожар на заводе (столь крупная ошибка не может явиться следствием того же нормального распределения, что и элемент случайности в ваших данных), непредвиденное изменение структуры затрат или дополнительные непредвиденные затраты, связанные с удвоением или утроением объема производства.



Рис. 11.2.12. Доверительный интервал для значения нового наблюдения Y при заданном значении X_0 зависит от того, насколько далеко X_0 отстоит от $\bar{X}$. Этот интервал оказывается наименьшим (что дает несколько большую точность) вблизи $\bar{X}$, где вы получаете наибольшую информацию от имеющихся у вас данных

Среднее значение Y: неопределенность и доверительный интервал

Если вас интересует *среднее значение* Y при заданном значении X_0, то для формирования доверительных интервалов и выполнения проверок гипотез вам понадобится соответствующая стандартная ошибка, которую обозначают $S_{(\text{прогнозируемое } Y|X_0)}$ (ее определение приведено ниже). Эти действия во многом напоминают те, что были описаны в предыдущем разделе, где величина $S_{(Y|X_0)}$ использовалась для статистического вывода о значении нового наблюдения Y при заданном значении X_0; правда, стандартная ошибка в данном случае вычисляется иначе.

Как сопоставить эти две ситуации (среднее значение Y при заданном X_0 и значение Y для одного наблюдения при заданном X_0)? В обоих случаях используется *одно и то же оценочное значение*, а именно прогнозируемое на основе линии наименьших квадратов значение $a + bX_0$. Однако, поскольку отдельные наблюдения имеют большую изменчивость, чем статистические характеристики, $S_{(Y|X_0)}$ больше, чем $S_{(\text{прогнозируемое } Y|X_0)}$. Это объясняется тем, что отдельное наблюдение Y ($\alpha + \beta X_0 + \varepsilon$ в соответствии с линейной моделью) включает и случайную ошибку, ε, тогда как среднее значение Y (равное $\alpha + \beta X_0$) — нет.

Это можно представить себе следующим образом. Проанализировав доходы (X) и расходы на спортивные товары (Y), вы можете составить неплохое представление о том, сколько в среднем тратит на приобретение спортивных товаров типичный покупатель, зарабатывающий $35 000 в год. Речь идет о *средней* сумме затрат на приобретение спортивных товаров для всех людей, зарабатывающих примерно $35 000 в год. Эту величину можно оценить довольно точно на большой выборке, поскольку это среднее значение, которое близко к средней сумме затрат всех людей из соответствующей генеральной совокупности, а именно тех, кто зарабатывает около $35 000 в год. Однако затраты *отдельных* покупателей могут существенно различаться — в конце концов, далеко не каждый играет в обеденный перерыв со своими клиентами в теннис. Вариация индивидуальных значений *не* усредняется даже в самой крупной выборке (здесь размер выборки n не имеет значения).

Если X равно известному значению, X_0, среднее значение для Y равняется $\alpha + \beta X_0$. Обратите внимание, что это среднее значение представляет собой параметр генеральной совокупности, поскольку оно является неизвестным и *фиксированным* (не случайным). Среднее значение для Y при заданном значении X_0 оценивается прогнозируемым значением $a + bX_0$, которое является *случайной* величиной, поскольку оценки a и b вычисляются методом наименьших квадратов на основе случайной выборки данных. Степень этой случайности характеризуется следующей формулой для стандартной ошибки прогнозируемого значения (среднего значения) Y при заданном значении X_0.

Стандартная ошибка прогнозируемого (среднего) значения Y при заданном значении X_0

$$S_{(\text{прогнозируемое } Y|X_0)} = \sqrt{S_e^2\left(\frac{1}{n}\right) + S_b^2(X_0 - \bar{X})^2}\ ;\ \text{число степеней свободы равно } n - 2.$$

Доверительный интервал для прогнозируемого (среднего) значения Y при заданном значении X_0

$$\text{От } (a + bX_0) - tS_{(\text{прогнозируемое } Y|X_0)} \text{ до } (a + bX_0) + tS_{(\text{прогнозируемое } Y|X_0)}.$$

Эта стандартная ошибка зависит от S_e (базовая неопределенность в данной ситуации), от S_b (неопределенность коэффициента регрессии, использованного для прогнозирования) и от расстояния между X_0 и $\bar{X}$. Эта ошибка будет меньше, когда X_0 оказывается ближе к $\bar{X}$, поскольку в этом случае вы располагаете большей информацией. Стандартная ошибка прогнозируемого (среднего) значения будет увеличиваться при удалении X_0 от $\bar{X}$, поскольку в этом случае имеющаяся информация (наблюдаемые значения X) недостаточно близка к информации, которая требуется (X_0). Это поведение проиллюстрировано на рис. 11.2.13 применительно к данным о производственных затратах.

Допустим, вы планируете в неопределенном будущем произвести $X_0 = 39$ изделий и вам требуется надежная оценка среднего значения недельных производственных затрат на длительную перспективу. Воспользовавшись уравнением прогноза, вы оценили эти затраты на уровне $a + bX_0 = 2272 + (51{,}66)(39) =$ $= \$4287$. Неопределенность в этой оценке недельных производственных затрат на длительную перспективу можно вычислить так:

$$S_{(\text{прогнозируемое } Y|X_0)} = \sqrt{S_e^2\left(\frac{1}{n}\right) + S_b^2(X_0 - \bar{X})^2} = \sqrt{198{,}58^2\left(\frac{1}{18}\right) + 7{,}35^2(39 - 32{,}50)^2} =$$

$$= \sqrt{(39{,}434)(0{,}055556) + (54{,}0225)(42{,}25)} = \sqrt{4473{,}25} = \$66{,}88.$$



Рис. 11.2.13. Доверительный интервал для прогнозируемого (среднего) значения Y при заданном значении X_0 зависит от расстояния между X_0 и $\bar{X}$. Этот интервал оказывается меньше доверительного интервала для отдельного наблюдения Y из-за отсутствия дополнительной случайности в значениях отдельных наблюдений (сравните с рис. 11.2.12)

Обратите внимание, насколько меньше эта стандартная ошибка ($66,88 для среднего значения) по сравнению со стандартной ошибкой для отдельной недели ($209,54), вычисленной нами в предыдущем разделе.[23]

95% доверительный интервал на основе t-значения (2,120) из t-таблицы при $n - 2 = 16$ степенях свободы простирается от $4287 – (2,120)(66,88) до $4287 + (2,120)(66,88). Иными словами:

"Мы на 95% уверены в том, что в долгосрочном плане средние недельные производственные затраты (прогноз для выпуска 39 изделий в неделю составляет $4287) окажутся в диапазоне от $4145 до $4429".

Такой доверительный интервал учитывает только статистическую ошибку в оценке прогнозируемого значения, $4287. Этот прогноз выполнен методом наименьших квадратов исходя из имеющейся относительно небольшой случайной выборки данных. Если линейная модель является адекватным описанием вашей структуры затрат, тогда такой доверительный интервал будет правильным. Однако, как и в предыдущем случае, этот статистический метод не позволяет учитывать другие, непредсказуемые источники ошибок.

Регрессия может вводить в заблуждение

Несмотря на то что регрессия является одним из наиболее мощных и полезных методов статистики, с применением этого метода связан ряд потенциальных проблем, о которых никогда не следует забывать. Поскольку вывод из регрессионного анализа основывается на линейной модели, полученные результаты могут оказаться неверными, если линейная модель не соответствует рассматриваемой

[23] Из-за ошибки округления мы теряем один цент. Если повысить точность S_b (использовать значение 7,3472 вместо 7,35), можно найти более точный ответ — $66,87.

генеральной совокупности. Ваш уровень ошибки может оказаться намного выше, чем предполагаемые 5%, ваш доверительный уровень может оказаться намного ниже предполагаемых 95% и, наконец, качество ваших прогнозов может оказаться намного хуже, чем они могли бы быть при отсутствии такой проблемы.

Поскольку вы располагаете весьма ограниченной совокупностью данных, вы имеете очень мало информации о наблюдениях, относительно которых ваши данные являются нерепрезентативными. Поскольку ваша регрессия основывается на наблюдаемой ситуации, она не может прогнозировать результаты некоторого вмешательства, порождающего новую ситуацию с новой динамикой. Более того, с чисто технической точки зрения, далеко не все равно, прогнозируете ли вы Y на основе X или, наоборот, X на основе Y.

Это лишь некоторые из проблем, о которых следует помнить, пользуясь результатами статистического анализа, выполненного самостоятельно или кем-либо другим. Ниже приведены дополнительные сведения, касающиеся ловушек, расставляемых регрессией.

Линейная модель может оказаться неверной

Вспомним, как выглядит линейная модель для генеральной совокупности:

$$Y = (\alpha + \beta X) + \varepsilon =$$
$$= (\text{взаимосвязь в генеральной совокупности}) + \text{случайность},$$

где ε имеет нормальное распределение со средним значением 0 и постоянным стандартным отклонением. Существует несколько причин, в силу которых эта взаимосвязь может не соответствовать генеральной совокупности.

Если истинная взаимосвязь *нелинейная*, вряд ли можно использовать полученную в результате оценки прямую линию для прогноза Y, как это показано на рис. 11.2.14 и 11.2.15. Большинство компьютерных программ не препятствует использованию метода наименьших квадратов в подобных ситуациях и только некоторые из них предупреждают о подобных проблемах. Вы сами должны анализировать данные и находить источники возможных проблем.

Процесс **экстраполяции** (а именно предсказания, выходящего за пределы диапазона значений данных, которые у вас имеются) особенно опасен, поскольку в таком случае вы не в состоянии защититься от возможных проблем путем анализа имеющихся данных. Подобная проблема проиллюстрирована на рис. 11.2.16.

Единственное резко отклоняющееся значение может коренным образом изменить ситуацию, как это показано на рис. 11.2.17. Предположение линейной модели о нормальном распределении фактора случайности подразумевает, что наличие значения, резко отклоняющегося от линии генеральной совокупности, чрезвычайно маловероятно. Линия наименьших квадратов пытается учесть и это значение, но это вступает в противоречие с ее способностью прогнозировать типичные случаи, в число которых резко отклоняющиеся значения просто не попадают. Решение этой проблемы возможно с помощью так называемых устойчивых регрессионных методов.[24]

[24] См., например, Hoaglin D. C., Mosteller F., and Tukey J. W. *Understanding Robust and Exploratory Data Analysis* (New York: Wiley, 1983).

Наконец, если для ваших данных характерна *непостоянная вариация*, ваши статистические выводы будут ненадежными. Слишком много внимания будет уделено части данных с высокой вариацией и слишком мало — данным с малой вариацией. Возможны два варианта решения этой проблемы: (1) преобразовать данные таким образом, чтобы уравнять вариацию и добиться их большего соответствия прямой линии; (2) воспользоваться более продвинутым методом *взвешенного* регрессионного анализа, чтобы сбалансировать важность различных наблюдений.

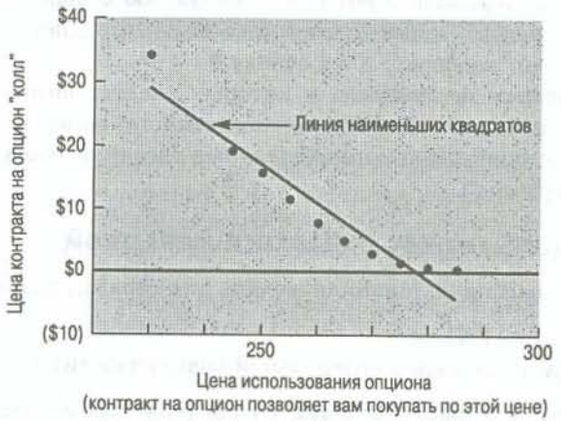


Рис. 11.2.14. Нелинейную взаимосвязь невозможно качественно прогнозировать с помощью прямой линии. Регрессия, базирующаяся на линейной модели, может предсказывать отрицательные цены опционных контрактов на основе фондового индекса при высоких ценах использования опциона, что совершенно невозможно с финансовой точки зрения

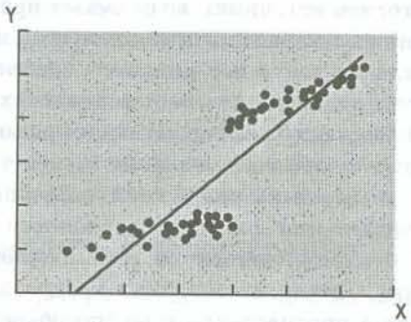


Рис. 11.2.15. Нелинейность может порождать так называемый "пороговый эффект", результатом которого также является плохое качество прогнозов. В данном случае это выглядит как наличие сильно различающихся групп (кластеринг). Результаты могут быть значительно лучше, если для каждой группы (кластера) построить свою линию регрессии

Теперь давайте посмотрим, какие неприятности с регрессией и ее интерпретацией подстерегают вас даже в тех случаях, когда анализируемая совокупность данных вполне *соответствует* линейной модели.

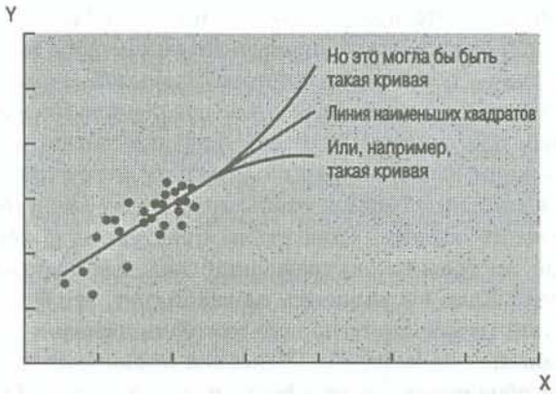


Рис. 11.2.16. Экстраполяция за пределы имеющихся данных потенциально опасна. Несмотря на то что генеральная совокупность может соответствовать прямой линии, вы все же располагаете недостаточной информацией, чтобы отбросить другие возможности. Две показанные на этом рисунке изогнутые линии очень похожи на прямые в том диапазоне данных, которыми вы располагаете. Однако, что находится за пределами этого диапазона, предсказать довольно трудно

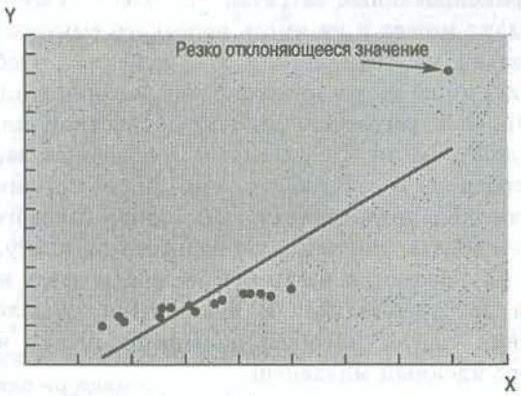


Рис. 11.2.17. Резко отклоняющееся значение может серьезно повлиять на результаты регрессионного анализа методом наименьших квадратов. Вашу способность делать качественные прогнозы в типичных случаях существенно ухудшает наличие одной-единственной резко отклоняющейся точки

Трудно предсказать вмешательство исходя из наблюдаемого опыта

Когда вы используете регрессию для прогнозирования на основе имеющихся данных, то предполагаете, что прогнозируемое новое наблюдение возникает из той же базовой системы, которая сгенерировала имеющиеся у вас данные. Если же меняется сама система — либо в результате собственной эволюции, либо вследствие вмешательства извне, — ваши прогнозы могут оказаться недействительными.

Можно, например, построить линию регрессии для прогнозирования объема новых заказов (в денежном выражении) на основании количества телефонных звонков, поступивших в магазин. Наклон этой линии будет свидетельствовать о средней "ценности" каждого звонка. Следует ли приступать к реализации той или иной маркетинговой программы, направленной на стимулирование подобных обращений по телефону? Если вы решитесь на такой шаг, это будет означать вмешательство в функционирование системы, что может привести к изменению структуры заказов, принимаемых по телефону. Подобная маркетинговая программа может генерировать новые обращения по телефону, целью которых будет получение дополнительной информации о товаре, но не желание немедленно его заказать. Не исключено, разумеется, что подобная кампания приведет к росту количества заказов; проблема заключается лишь в том, что вычисленный вами наклон (основанный на предыдущих данных) может не отражать поведение новой системы.

Сдвиг может быть лишен смысла

Построив линию регрессии, связывающую данные о затратах (Y) с количеством произведенной продукции (X), сдвиг (отрезок, отсекаемый на оси Y) мы интерпретируем как фиксированные затраты, что очень важно для нас. Однако в других ситуациях сдвиг может и не иметь полезного смысла. Он может быть необходим исключительно из технических соображений, чтобы получить оптимальное предсказание, но не иметь практической интерпретации.

Рассмотрим, например, регрессию размера заработной платы (Y) на возраст работника (X). Наклон этой линии указывает добавочную заработную плату, на которую можно (в среднем) рассчитывать, имея более старший возраст. Некоторый сдвиг необходим для установления некоторого базового уровня, что дает возможность прогнозировать фактическую заработную плату, например с помощью уравнения $a + bX$. Несмотря на то что без a нам здесь не обойтись, его интерпретация весьма затруднительна. В буквальном смысле он соответствует ожидаемой заработной плате, которую должен получать человек в возрасте $X = 0$, то есть новорожденный младенец!

Вообще говоря, это не представляет собой проблемы. В случаях, подобных описанному выше, об интерпретации этого отрезка можно вообще не задумываться.[25]

[25] Одним из способов решения этой проблемы является использование вместо сдвига так называемого центрального значения. Линия в этом случае описывается уравнением $Y = c + b(X - \bar{X})$. Центральное значение c представляет собой ожидаемое значение Y для наиболее типичного значения X, а именно для $\bar{X}$, поэтому интерпретация в таком случае не представляет сложности. Наклон имеет тот же смысл, что и ранее.

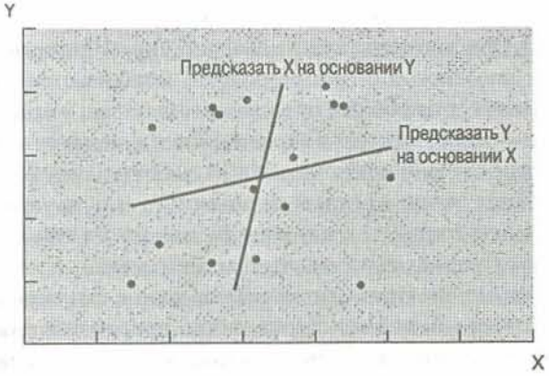


Рис. 11.2.18. Эти две линии регрессии соответствуют ситуациям, когда Y прогнозируется на основании X (обычная процедура) и когда X прогнозируется на основании Y. Поскольку в данной ситуации налицо значительный фактор случайности, линии сильно отличаются друг от друга. Каждая из этих линий достаточно хорошо прогнозирует среднее значение ($\bar{X}$ или $\bar{Y}$) соответствующей переменной (т.е. горизонтальной или вертикальной оси)

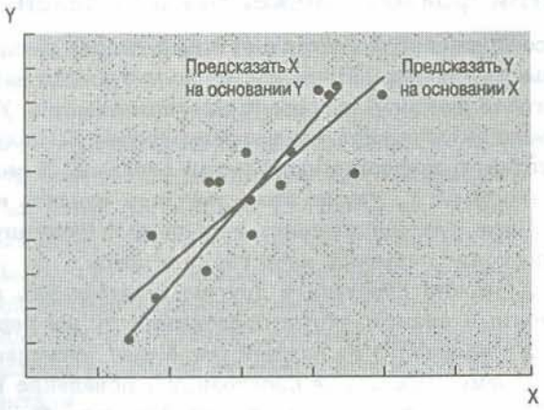


Рис. 11.2.19. Две линии регрессии сближаются, когда уменьшается фактор случайности и точки данных приближаются к прямой линии. Когда точки данных окажутся точно на прямой линии, две линии регрессии совпадут

Представление *Y* на основании *X* и представление *X* на основании *Y*

Какая именно из двух переменных прогнозируется, а какая служит основанием для прогноза, действительно *имеет большое значение*: прогнозирование Y на основании X отличается от прогнозирования X на основании Y, причем каждому из

этих подходов соответствует своя линия регрессии. Это вполне объяснимо, так как каждому из этих случаев соответствуют свои ошибки. Например, прогнозирование производительности на основании стажа работы связано с ошибками прогнозирования, выражающимися в единицах производительности, тогда как прогнозирование стажа работы на основании производительности связано с ошибками прогнозирования, выражающимися в единицах стажа работы. Разумеется, если все ваши точки данных попадают точно на прямую линию (в результате чего коэффициент корреляции будет равен 1 или −1), эту линию можно использовать для прогнозирования любой из двух переменных на основании другой.

Однако в обычном случае приходится иметь дело с фактором случайности или неопределенности, который подталкивает ваши прогнозируемые значения в направлении среднего значения той переменной, которая прогнозируется (X или Y). В экстремальном случае, когда мы имеем дело с чистой случайностью, наилучшим прогнозом Y на основании X является $\bar{Y}$, а наилучшим прогнозом X — Y является $\bar{X}$. Помните формулу наклона $b = rS_Y/S_X$? Она указывает на то, что линия становится более пологой (менее крутой), когда возникает большая неопределенность (корреляция, r, приближается к 0).

На рис. 11.2.18 и 11.2.19 показаны две линии регрессии. Обратите внимание: когда точки данных оказываются ближе к линии, линии регрессии также сближаются, поскольку линия в этом случае лучше определяется данными.

Скрытый "третий фактор" может быть полезен

Это последнее соображение представляет собой скорее не проблему, а некоторое предложение по улучшению. Несмотря на то что линия наименьших квадратов представляет собой наилучший способ прогнозирования Y на основании X, всегда есть возможность улучшить качество прогнозов Y, получив в свое распоряжение дополнительную информацию. Иными словами, X может не содержать достаточно информации об Y, что не позволяет нам сделать качественный прогноз Y; возможно, вам удастся выявить еще одну переменную (некий третий фактор), которая позволит повысить качество прогнозов.

Если на место X можно подставить другую переменную, можно выполнить еще один регрессионный анализ, чтобы предсказать ту же переменную Y. Сравнение членов $R2$ (или членов S_e) из каждой регрессии показывает, какая из этих двух поясняющих переменных лучше прогнозирует поведение Y.

Если вы хотите объединить информацию из *двух или больше X-переменных*, то вам следует воспользоваться *множественной регрессией* — чрезвычайно важным методом, применяемым в бизнесе и исследованиях. Этот метод рассматривается в следующей главе.

11.3. Дополнительный материал

Резюме

Тремя основными целями анализа двумерных данных, представленных парами (X,Y), являются (1) описание и понимание взаимосвязи, (2) прогнозирование и предсказание нового наблюдения и (3) корректировка и управление процессом.

Корреляционный анализ позволяет сделать вывод о силе взаимосвязи, а *регрессионный анализ* используется для прогнозирования одной переменной на основании другой (как правило, Y на основании X).

Двумерные данные анализируют с использованием **диаграммы рассеяния** в координатах Y и X, которая дает визуальное представление взаимосвязи в данных. **Корреляция**, или **коэффициент корреляции** (r), представляет собой безразмерное (не имеющее единиц измерения) число в диапазоне от −1 до 1, которое характеризует силу взаимосвязи. Равенство коэффициента корреляции 1 свидетельствует об идеальной взаимосвязи в виде прямой линии с наклоном вверх. Равенство коэффициента корреляции −1 свидетельствует об идеальной взаимосвязи в виде наклоненной вниз (отрицательно) прямой линии. Коэффициент корреляции говорит о том, насколько близко к этой наклоненной прямой линии расположены точки диаграммы, однако он не характеризует крутизну наклона этой линии. Формула вычисления коэффициента корреляции имеет следующий вид:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}.$$

Ковариация X и Y представляет собой числитель в формуле для коэффициента корреляции. Поскольку единицы измерения ковариации трудно интерпретировать, удобнее работать с коэффициентом корреляции.

При анализе двумерной диаграммы рассеяния можно обнаружить различные взаимосвязи. Простейшей, с точки зрения анализа, является **линейная взаимосвязь**, которая выражается в том, что точки на диаграмме рассеяния с постоянным разбросом группируются случайным образом вдоль прямой линии. Диаграмма свидетельствует об **отсутствии взаимосвязи**, если точки размещены случайно и при перемещении слева направо невозможно обнаружить какой-либо уклон (ни вверх, ни вниз). Двумерная диаграмма рассеяния характеризуется **нелинейной взаимосвязью**, если точки на ней группируются вдоль *кривой*, а не прямой линии. Поскольку количество видов кривых практически безгранично, анализ нелинейной взаимосвязи оказывается намного сложнее, однако взаимосвязь можно приблизить к линейной, применив к данным соответствующее преобразование. Проблема **неравной вариации** возникает тогда, когда при перемещении по горизонтали на диаграмме рассеяния вариация точек по вертикали сильно меняется. Неравная вариация приводит к снижению надежности коэффициента корреляции и регрессионного анализа. Проблему неравной вариации можно решить с помощью соответствующих преобразований данных или с помощью так называемой взвешенной регрессии. Проблема **кластеринга** (разделение совокупности на группы более однородных объектов) возникает в случае образования на диаграмме рассеяния отдельных, ярко выраженных групп точек; в таких случаях каждую группу следует анализировать отдельно. Некоторая точка данных является **выбросом (резко отклоняющимся значением)**, если она не соответствует взаимосвязи между остальными данными; резко отклоняющиеся значения могут исказить статистические характеристики двумерной совокупности данных.

Корреляцию нельзя рассматривать как причинную обусловленность. Коэффициент корреляции характеризует связь между числами, но не объясняет ее. Корреляция может быть вызвана тем, что переменная X влияет на Y, или тем, что переменная Y влияет на X. Кроме того, корреляция может быть вызвана также тем, что на X и Y влияет некий скрытый "третий фактор", что создает впечатление связи между X и Y. Термином **ложная корреляция** обозначают высокую корреляцию, которая возникает благодаря действию некоторого третьего фактора.

Регрессионный анализ заключается в прогнозировании одной переменной на основании другой. **Линейный регрессионный анализ** прогнозирует значение одной переменной на основании другой с помощью прямой линии. **Наклон** этой линии, b, выражается в единицах измерения Y на одну единицу X и характеризует крутизну подъема или спуска (если b отрицательное) линии. **Сдвиг**, a, равен значению, которое принимает Y при X, равном 0. Уравнение прямой линии имеет следующий вид:

$$Y = Сдвиг + (Наклон)(X) = a + bX .$$

Линия наименьших квадратов характеризуется наименьшей из всех возможных линий суммой возведенных в квадрат ошибок прогнозирования по вертикали и используется как лучшая линия прогнозирования, основанная на данных. Наклон b называют также **коэффициентом регрессии** Y по X, а сдвиг a (отрезок, отсекаемый на оси Y) называют также **постоянным членом** регрессии. Ниже приведены уравнения для наклона и сдвига, соответствующие линии наименьших квадратов.

$$\text{Наклон равен: } b = r\frac{S_Y}{S_X} .$$

$$\text{Сдвиг равен: } a = \bar{Y} - b\bar{X} = \bar{Y} - r\frac{S_Y}{S_X}\bar{X} .$$

Формула для линии наименьших квадратов имеет следующий вид:

$$\text{(Прогнозируемое значение } Y) = a + bX = (\bar{Y} - r\frac{S_Y}{S_X}\bar{X}) + r\frac{S_Y}{S_X}X .$$

Прогнозируемое значение для Y при заданном значении X определяется путем подстановки этого значения X в уравнение для линии наименьших квадратов. Каждая из точек данных характеризуется **остатком** — ошибкой прогнозирования, указывающей, насколько выше или ниже линии находится точка.

Существуют две меры соответствия линии наименьших квадратов имеющимся данным. **Стандартная ошибка оценки**, которую обозначают S_e, приблизительно указывает величину ошибок прогнозирования (остатков) для имеющихся данных *в тех же единицах, в которых измерена переменная* Y. Соответствующие формулы приведены ниже.

$$S_e = S_Y\sqrt{(1-r^2)\frac{n-1}{n-2}} \quad \text{(для вычисления)}$$

$$= \sqrt{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n}[Y_i - (a+bX_i)]^2} \quad \text{(для интерпретации).}$$

Значение $R2$, часто называемое **коэффициентом детерминации**, говорит о том, какой процент вариации Y объясняется поведением X.

Доверительные интервалы и проверка гипотез для коэффициента регрессии связаны с определенными предположениями относительно анализируемой совокупности данных, которые должны гарантировать, что она состоит из независимых наблюдений, характеризующихся линейной взаимосвязью с равной вариацией и приблизительно нормально распределенной случайностью. Во-первых, эти данные должны представлять собой произвольную выборку из интересующей нас генеральной совокупности. Во-вторых, **линейная модель** указывает, что наблюдаемое значение Y определяется взаимосвязью в генеральной совокупности плюс случайная ошибка, имеющая нормальное распределение. Существуют параметры генеральной совокупности, соответствующие наклону и сдвигу линии наименьших квадратов, построенной на данных выборки:

$$Y = (\alpha + \beta X) + \varepsilon =$$
$$= \text{(Взаимосвязь в генеральной совокупности)} + \text{случайность.}$$

где ε имеет нормальное распределение со средним значением, равным 0, и постоянным стандартным отклонением σ.

Статистические выводы (использование доверительных интервалов и проверки статистических гипотез) относительно коэффициентов линии наименьших квадратов основываются, как обычно, на их стандартных ошибках и значениях из t-таблицы для $n - 2$ степеней свободы. **Стандартная ошибка коэффициента наклона**, S_b, указывает приблизительную величину отклонения оценки наклона, b (коэффициент регрессии, вычисленный на основе данных выборки), от наклона в генеральной совокупности, β, вызванного случайным характером выборки. **Стандартная ошибка сдвига**, Sa, указывает приблизительно, насколько далеко оценка сдвига a отстоит от истинного сдвига α в генеральной совокупности. Соответствующие формулы выглядят следующим образом:

стандартная ошибка коэффициента регрессии:

$$S_b = \frac{S_e}{S_X\sqrt{n-1}};$$

стандартная ошибка сдвига:

$$S_a = S_e\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_X^2(n-1)}} .$$

Доверительный интервал для наклона в генеральной совокупности, β:

$$\text{от } b - tS_b \text{ до } b + tS_b.$$

Доверительный интервал для сдвига в генеральной совокупности, α:

$$\text{от } a - tS_a \text{ до } a + tS_a.$$

Один из способов проверки, является ли обнаруженная взаимосвязь между X и Y реальной или это просто случайное совпадение, заключается в сравнении β с заданным значением $\beta_0 = 0$. О значимой связи можно говорить в том случае, ес-

ли 0 не попадает в доверительный интервал, базирующийся на b и S_b, или если абсолютное значение $t = b/S_b$ превосходит соответствующее t-значение в t-таблице. Эта проверка эквивалентна проверке значимости коэффициента корреляции и означает, по сути, то же самое, что и F-тест во множественной регрессии (см. следующую главу) для случая, когда уравнение содержит только одну переменную X. Разумеется, любой из коэффициентов (a или b) можно сравнить с любым подходящим заданным значением, воспользовавшись одно- или двусторонней проверкой (в зависимости от конкретных обстоятельств) и с использованием тех же методов проверки, что были рассмотрены в главе 10 для среднего генеральной совокупности.

Для прогнозирования значения нового наблюдения Y при условии, что $X = X_0$, неопределенность прогноза оценивают с помощью стандартной ошибки $S_{(Y|X_0)}$, которая также имеет $n - 2$ степеней свободы. Это позволяет построить доверительные интервалы и проверить гипотезы для нового наблюдения. Другая формула позволяет вычислить стандартную ошибку для прогнозирования *среднего* значения Y при заданном X_0:

$$S_{(Y|X_0)} = \sqrt{S_e^2\left(1+\frac{1}{n}\right)+S_b^2(X_0-\bar{X})^2}\ .$$

Доверительный интервал для нового наблюдения Y при заданном значении X_0 имеет следующий вид:

$$\text{от } (a+bX_0)-tS_{(Y|X_0)} \text{ до } (a+bX_0)+tS_{(Y|X_0)}\ .$$

С регрессионным анализом связаны определенные проблемы. Если линейная модель неадекватно описывает генеральную совокупность, прогнозы и статистические выводы будут недостоверны. Анализ диаграммы рассеяния позволяет выявить такие проблемы, как нелинейность, неравная вариация и наличие выбросов, что может помочь определить адекватность линейной модели. Процесс **экстраполяции**, позволяющий сделать прогноз за пределы диапазона значений имеющихся данных, особенно ненадежен, поскольку вы не в состоянии подстраховать себя путем анализа диаграммы рассеяния.

Даже если линейная модель соответствует изучаемой генеральной совокупности, проблемы все равно остаются. Поскольку прогнозы регрессии базируются на данных из прошлого, регрессия не в состоянии идеально прогнозировать последствия вмешательства, которое изменяет структуру самой системы. В некоторых случаях бывает трудно интерпретировать сдвиг a, хотя этот член и является неотъемлемой частью уравнения прогноза, построенного методом наименьших квадратов. Обратите особое внимание на выбор переменной, которую вы будете прогнозировать, поскольку линия прогнозирования Y на основании X отличается от линии прогнозирования X на основании Y (особенно при наличии существенной случайности в данных). Наконец, может существовать и некий третий фактор, который способен повысить качество прогнозирования Y по сравнению с использованием одной лишь переменной X; эту ситуацию мы обсудим в следующей главе.

Основные термины

- Диаграмма рассеяния (scatterplot), 520
- Коэффициент корреляции (correlation coefficient), r, 521
- Ковариация (covariance), 524
- Линейная взаимосвязь (linear relationship), 525
- Отсутствие взаимосвязи (no relationship), 531
- Нелинейная взаимосвязь (nonlinear relationship), 533
- Неравная вариация (unequal variability), 537
- Кластеринг, разделение совокупности на относительно однородные группы (clustering), 540
- Выброс (резко отклоняющееся значение) в двумерных данных (bivariate outlier), 542
- Ложная корреляция (spurious correlation), 545
- Линейный регрессионный анализ (linear regression analysis), 547
- Наклон (slope), 549
- Сдвиг (intercept), 549
- Линия наименьших квадратов (least-squares line), $Y = a + bX$, 549
- Коэффициент регрессии (regression coefficient), b, 549
- Постоянный член (constant term), a, 549
- Прогнозируемое значение (predicted value), 551
- Остаток (residual), 551
- Стандартная ошибка оценки (standard error of estimate), S_e, 555
- Коэффициент детерминации (coefficient of determination), R^2, 558
- Линейная модель (linear model), 559
- Стандартная ошибка коэффициента наклона (standard error of the slope coefficient), S_b, 560
- Стандартная ошибка сдвига (standard error of the intercept term), S_a, 561
- Экстраполяция (extrapolation), 573

Контрольные вопросы

1. В чем отличие анализа двумерных данных от анализа одномерных данных?
2. В чем разница между корреляционным анализом и регрессионным анализом?
3. Какой вид анализа (корреляционный или регрессионный) применяется в каждой из описанных ниже ситуаций?

   а) Выяснение наличия какой-либо взаимосвязи между расходами на рекламу и объемом продаж.

   б) Разработка системы прогнозирования эффективности портфеля ценных бумаг, основанной на изменениях одного из ведущих индексов фондовой биржи.

в) Создание инструмента формирования сметы, позволяющего выражать затраты в терминах количества произведенных изделий.

г) Анализ данных с целью определения силы взаимосвязи между моральным состоянием работников и их производительностью.

4. Для каждого из приведенных ниже равенств укажите типичный вариант интерпретации. Затем укажите, есть ли какие-то другие варианты.

а) $r = 1$.

б) $r = 0{,}85$.

в) $r = 0$.

г) $r = -0{,}15$.

д) $r = -1$.

5. а) Что представляет собой ковариация между X и Y?

б) Что легче интерпретировать, ковариацию или корреляцию? Почему?

6. Постройте диаграмму рассеяния, которая иллюстрировала бы каждый из перечисленных ниже типов структуры двумерных данных. Для ответа на этот вопрос не обязательно использовать какие-либо данные, можно непосредственно рисовать точки.

а) Взаимосвязь между X и Y отсутствует.

б) Линейная взаимосвязь с сильной положительной корреляцией.

в) Линейная взаимосвязь со слабой отрицательной корреляцией.

г) Линейная взаимосвязь с корреляцией -1.

д) Положительная связь с неравной вариацией.

е) Нелинейная взаимосвязь.

ж) Кластеринг (разделение совокупности данных на относительно однородные группы).

з) Положительная связь с резко отклоняющимся значением.

7. а) Если большие значения X вызывают появление больших значений Y, то какой, по вашему мнению, должна быть корреляция — положительной, отрицательной или нулевой? Почему?

б) Если вы выявили сильную положительную корреляцию, говорит ли это о том, что большие значения X вызывают появление больших значений Y? Если нет, какие еще возможны варианты?

8. а) Чем именно линия наименьших квадратов так отличается от всех других линий?

б) Откуда линии наименьших квадратов "известно", что она прогнозирует Y на основании X, а не наоборот?

в) Есть все основания считать "наиболее типичным" значением данных такое, у которого значение X равно $\bar{X}$, а значение Y — $\bar{Y}$. Покажите, что линия наименьших квадратов проходит через эту "наиболее типичную" точку.

г) Допустим, что стандартные отклонения X и Y остаются неизменными, тогда как корреляция уменьшилась, оставшись при этом положительной. Что в таком случае происходит с коэффициентом наклона b?

9. Дайте определение прогнозируемого значения и остатка для некоторой точки данных.

10. Для каждой из описанных ниже ситуаций укажите, какое из двух чисел, прогнозируемое значение или остаток, лучше использовать.

а) При составлении сметы необходимо знать, какое число следует указать в графе "затраты на проданные товары", исходя из ожидаемого объема продаж на следующий квартал.

б) Вам хотелось бы знать, насколько успешно работают подчиненные вам подразделения с учетом того, какими, по вашему мнению, должны были бы оказаться результаты их деятельности при наличии всех необходимых им ресурсов.

в) Чтобы определить размер заработной платы нового сотрудника, вы хотите знать разумный размер заработка сотрудника с таким же стажем работы.

г) В отчете, посвященном анализу кадровой политики своего предприятия, вы хотели бы отобразить, насколько больше (или меньше) получает каждый работник в сравнении с ожидаемой заработной платой для работника с таким же стажем.

11. В чем разница между стандартной ошибкой оценки и коэффициентом детерминации?

12. а) Какое значение R^2 лучше, более низкое или более высокое?

б) Какое значение S_e лучше, более низкое или более высокое?

13. а) Что такое линейная модель?

б) Какие два параметра определяют взаимосвязь в генеральной совокупности в виде прямой линии?

в) Какие выборочные статистики используются для оценки трех параметров генеральной совокупности: α, β или σ?

г) Наклон линии наименьших квадратов, построенной на основании данных выборки, — это параметр или статистика? Из чего это следует?

14. Дайте определение и напишите формулу вычисления для каждой из перечисленных ниже величин, которые используются в статистических выводах.

а) Стандартная ошибка, используемая для коэффициента регрессии.

б) Стандартная ошибка сдвига.

в) Стандартная ошибка нового наблюдения.

г) Число степеней свободы для каждой из этих стандартных ошибок.

15. Статистический вывод в регрессии основывается на линейной модели. Назовите по меньшей мере три проблемы, возникающие в случае несоответствия данных линейной модели.

16. Что такое экстраполяция? В чем заключается особая потенциальная опасность ее применения?

17. С помощью линии наименьших квадратов вы предсказали, что, исходя из ожидаемого объема продаж ($38 200 000), себестоимость реализованной продукции в конце следующего квартала поднимется до $8 330 000. Ваш приятель из соседнего офиса замечает: "Не означает ли это и то, что себестоимость реализованной продукции, равная $8 330 000, приводит к ожидаемому объему продаж на уровне $38 200 000?" Справедлив ли такой вывод? Поясните свой ответ. (*Подсказка*. Где в каждом из этих случаев X, а где Y? То есть, что предсказывается и на основании чего?)

18. а) Приведите пример, в котором сдвиг a имеет естественную интерпретацию.

б) Приведите пример, в котором сдвиг a не имеет естественной интерпретации.

Задачи

1. Рассмотрим совокупность данных из табл. 11.3.1, представляющую срок службы (в годах) и затраты на техническое обслуживание (в тысячах долларов за год) для пяти одинаковых печатных прессов.



а) Постройте диаграмму рассеяния для этой совокупности данных. Какому типу взаимосвязи соответствует эта диаграмма?

б) Вычислите корреляцию между сроком службы и затратами на техническое обслуживание. Какой вывод можно сделать на основании этой корреляции?

в) С помощью метода наименьших квадратов найдите уравнение регрессии, которое позволяло бы прогнозировать затраты на техобслуживание на основании срока службы оборудования. Изобразите соответствующую линию на диаграмме рассеяния.

г) Какими, по вашему мнению, будут годовые затраты на техобслуживание одного пресса с семилетним сроком службы?

д) Какова в этом случае типичная величина ошибок прогнозирования?

е) Какая часть вариации затрат на техобслуживание объясняется тем, что срок службы одних прессов больше, чем других?

ж) Объясняет ли срок службы оборудования значимую часть вариации затрат на техобслуживание? Как вы об этом узнали?

з) Ваш консервативно настроенный заместитель предложил запланировать добавочные годовые затраты на техобслуживание в объеме $20 000 на один год возраста каждого пресса. Выполните проверку гипотез на уровне

Таблица 11.3.1

Срок службы	Затраты на техобслуживание
2	6
5	13
9	23
3	5
8	22

5% и выясните, значимо ли добавочные годовые затраты отличаются от тех, которые предлагает ваш заместитель.

2. Анализ линейной регрессии привел к следующему уравнению, связывающему доход с количеством часов, затраченным руководством фирмы на разработку проектов в прошлом году:

$$\text{доход} = -\$957 + \$85 \times \text{количество часов}.$$

а) В соответствии с этой оценкой взаимосвязи укажите, каким был бы доход (или убытки), если бы на планирование вообще не тратилось время?

б) Насколько в среднем увеличиваются доходы от проектов при увеличении затраченного на планирование времени на 10 часов?

в) Найдите точку самоокупаемости, представляющую собой количество часов, при которых оцениваемая величина дохода равна нулю.

г) Если корреляция r равна 0,351, какой процент вариации дохода объясняется временем, затрачиваемым на планирование?

д) Какая часть вариации дохода не объясняется количеством времени, затраченным на планирование? Напишите небольшой текст, поясняющий, насколько можно доверять этому уравнению прогнозирования, и обсудите другие факторы, которые могут оказывать влияние на уровень дохода.

3. В табл. 11.3.2 представлены показатели работы 10 крупнейших авиалиний за один месяц (октябрь 1995 г.) и за целый год, заканчивающийся в октябре 1995 г. Представленные в таблице числа показывают процент авиарейсов, прибывших без задержек. Мы проанализируем стабильность работы авиалиний, выявив взаимосвязь между показателями за один месяц и за целый год (если таковая взаимосвязь вообще существует).

а) Постройте диаграмму рассеяния точек для этой совокупности данных и выскажите свои соображения по поводу наличия взаимосвязи в этих данных.

б) Определите корреляцию между показателями работы авиалиний за один месяц и за целый год. Прослеживается ли в этом случае сильная взаимосвязь?

в) Определите коэффициент детерминации и скажите, что он представляет.

г) Определите уравнение линейной регрессии, позволяющее прогнозировать показатели работы авиалиний за один месяц, исходя из показателей за целый год включая и этот месяц.

д) Определите прогнозируемое значение и остаток для Alaska Airlines. Что представляет каждое из этих значений?

е) Определите стандартную ошибку оценки. Что она измеряет?

ж) Определите стандартную ошибку коэффициента регрессии.

з) Определите 95% доверительный интервал для коэффициента регрессии.

и) Выполните проверку гипотез на уровне 5% и выясните, существует ли значимая взаимосвязь между показателями работы авиалиний за эти два периода времени. Говорит ли это что-нибудь о стабильности работы авиалиний?

4. Закрытые инвестиционные фонды — в отличие от обычных паевых инвестиционных фондов открытого типа, которые непрерывно занимаются покупкой и продажей акций, — продают свои акции в виде фиксированной

"корзины" (портфеля) ценных бумаг. Рассмотрим стоимость чистых активов фонда и биржевую цену для World Income Funds, представленных в табл. 11.3.3. Несмотря на то что можно было бы ожидать, что каждый фонд будет продавать свои акции (биржевая цена) по той же цене, что и сумма его компонентов (стоимость чистых активов фонда в расчете на одну акцию), как правило, здесь наблюдается определенное несоответствие.

а) Насколько сильна взаимосвязь между стоимостью чистых активов фонда в расчете на одну акцию и биржевой ценой акции для этих инвестиционных фондов закрытого типа?

б) Можно ли считать существенной взаимосвязь между стоимостью чистых активов фонда в расчете на одну акцию и биржевой ценой акций или все

Таблица 11.3.2. Показатели работы авиалиний (соблюдение расписания рейсов)

Авиалиния	Один месяц	Весь год	Авиалиния	Один месяц	Весь год
Southwest	87,0	82,5	Alaska Air	81,4	76,3
Northwest	86,0	81,9	TWA	81,0	75,1
Continental	85,3	79,6	America West	80,5	78,0
American	84,7	78,2	United	79,2	80,0
USAir	82,4	81,1	Delta	75,2	78,7

Данные взяты из "How to Make an Airline Run on Schedule" by Scott McCartney, *The Wall Street Journal*, 1995, December 22, p. B1. Источник данных: U.S. Department of Transportation Air Travel Consumer Report, October 1995.

Таблица 11.3.3. Международные инвестиционные фонды закрытого типа

Фонд	Стоимость чистых активов фонда в расчете на одну акцию, дол.	Биржевая цена, дол.
ACM Mgd Multi-Mkt	9,50	8,750
Blackrock No Am	12,72	12,500
Emerging Mkts Inc	13,89	14,250
First Australia Pr	9,44	10,1875
First Commonwealth	12,54	11,875
Global Government	7,38	7,000
Global Income Plus	9,18	8,750
Global Yield	8,10	7,500
Kleinwort Ben Aust	9,84	9,125
Lat Am Dollar Inc	12,95	12,875
Strategic Global Inc	13,47	12,125
Templeton GI Govt	8,19	8,625
Templeton GI Inc	8,17	9,000

Данные взяты из *The Wall Street Journal*, 1993, January 4, p. C13.

выглядит так, будто биржевые цены назначаются фондам случайным образом? Поясните свой ответ.

в) Определите линию наименьших квадратов, позволяющую прогнозировать биржевую цену, исходя из стоимости чистых активов фонда в расчете на одну акцию.

г) Значительно ли отличается от 1 наклон линии наименьших квадратов? Интерпретируйте свой ответ с точки зрения следующего вопроса: "Может ли быть так, что увеличение стоимости чистых активов фонда на один пункт приводит в среднем к увеличению биржевой цены тоже на один пункт?"

5. Рассмотрим количество сделок и общий объем денег, связанных с деятельностью по слиянию и приобретению компаний ведущими инвестиционными банками (см. табл. 11.1.4).

а) Определите уравнение регрессии, позволяющее прогнозировать объем денег на основании количества сделок.

б) Какой будет для этих инвестиционных банков оценка суммы отдельной дополнительной сделки?

в) Изобразите диаграмму рассеяния и линию регрессии для этой совокупности данных.

г) Найдите ожидаемую сумму сделок для Goldman Sachs и значение остатка. Интерпретируйте оба этих значения с точки зрения бизнеса.

д) Определите стандартную ошибку коэффициента наклона. Что означает это число?

е) Найдите 95% доверительный интервал для ожидаемой предельной стоимости дополнительной сделки для этих фирм. (Так на языке экономистов называется наклон.)

ж) Проверьте на уровне 5%, существует ли значимая взаимосвязь между количеством сделок и их общей стоимостью.

з) Ваш инвестиционный банк намеревается в следующем году войти в лидирующую группу подобных учреждений, приняв участие в 100 сделках. Допуская, что вам удастся добиться поставленной цели, вычислите 95% доверительный интервал для той суммы, которой вы будете оперировать.

6. Рассмотрим весьма щекотливую тему экономических банкротств. В табл. 11.3.4 для каждого штата указаны количество банкротств и численность населения штата (в тысячах).

а) Постройте диаграмму рассеяния, отображающую зависимость количества банкротств (Y) от численности населения (X). Опишите взаимосвязь, которую вам удалось (если удалось) выявить. Прослеживается ли здесь какая-то связь?

б) Есть ли соответствие линейной модели? Поясните свой ответ.

в) Найдите логарифм каждого значения (как численности населения, так и количества банкротств). Можно использовать либо десятичные, либо натуральные логарифмы, но только одного типа.

Таблица 11.3.4. Количество банкротств в различных штатах

Штат	Количество банкротств	Численность населения	Штат	Количество банкротств	Численность населения
Алабама	841	4187	Миссури	1230	5234
Аляска	108	599	Монтана	173	839
Аризона	2964	3936	Небраска	399	1607
Арканзас	186	2424	Невада	568	1389
Калифорния	19 695	31 211	Нью-Хэмпшир	617	1125
Колорадо	1542	3566	Нью-Джерси	2843	7879
Коннектикут	1093	3277	Нью-Мексико	448	1616
Делавэр	137	700	Нью-Йорк	6916	18 197
Округ Колумбия	200	578	Северная Каролина	1194	6945
Флорида	5088	13 679	Северная Дакота	145	635
Джорджия	2350	6917	Огайо	2127	11 091
Гавайи	305	1172	Оклахома	1440	3231
Айдахо	350	1099	Орегон	969	3032
Иллинойс	2094	11 697	Пенсильвания	3124	12 048
Индиана	1091	5713	Род-Айленд	344	1000
Айова	507	2814	Южная Каролина	392	3643
Канзас	1069	2531	Южная Дакота	175	715
Кентукки	841	3789	Теннесси	1209	5099
Луизиана	664	4295	Техас	7096	18 031
Мэн	383	1239	Юта	351	1860
Мэриленд	1540	4965	Вермонт	173	576
Массачусетс	2720	6012	Виргиния	1738	6491
Мичиган	2546	9478	Вашингтон	2025	5255
Миннесота	921	4517	Западная Виргиния	315	1820
Миссисипи	322	2643	Висконсин	1224	5038
			Вайоминг	90	470

Данные взяты из материалов Бюро переписи населения США (U.S. Bureau of the Census), *Statistical Abstract of the United States: 1994*, 114th edition (Washington, D.C., 1994), p. 27, 547. Численность населения указана в тысячах человек. Данные за 1993 г.

г) Постройте диаграмму рассеяния для логарифмов и опишите выявленную взаимосвязь.

д) Найдите уравнение линии регрессии для прогнозирования логарифма количества банкротств на основании логарифма численности населения.

е) Найдите двусторонний 95% доверительный интервал для коэффициента наклона взаимосвязи логарифмов значений.

ж) Проверьте на уровне 5%, существует ли значимая взаимосвязь между логарифмами количества банкротств и численности населения. Объясните, почему полученный результат является вполне разумным?

з) Если бы коэффициент наклона для логарифмов равнялся точно 1, тогда количество банкротств было бы пропорционально численности населения. Значение, большее 1, свидетельствовало бы о том, что в более крупных штатах наблюдается пропорционально большее количество банкротств, а наклон, меньший 1, указывал бы на то, что пропорционально большее количество банкротств наблюдается в меньших штатах. Проверьте на уровне 5%, значимо ли отличается наклон для логарифмов данных от 1, и сделайте краткие выводы.

7. В табл. 11.3.5 представлены суточные процентные изменения курса акций McDonald's и индекса Dow Jones Industrial Average для торговых дней в ноябре и декабре 1998 г.

а) Постройте диаграмму рассеяния, отображающую взаимосвязь между суточными процентными изменениями курса акций McDonald's и индекса Dow Jones.

б) Опишите взаимосвязь, которую вам удастся выявить на этой диаграмме.

в) Определите корреляцию между этими процентными изменениями. Соответствует ли найденное значение корреляции вашему впечатлению от диаграммы рассеяния?

г) Определите коэффициент детерминации. (Можно просто возвести в квадрат коэффициент корреляции.) Интерпретируйте эту величину как "объясненную вариацию". На языке финансистов этот показатель отражает долю *недиверсифицируемого риска* в McDonald's. Если, например, он равняется 100%, курс акций McDonald's идеально отслеживает поведение биржи, а диверсификация не дает ничего нового.

д) Определите долю *диверсифицируемого риска*. Она равняется $1 - R2$ (или 100% минус процент недиверсифицируемого риска). Доля диверсифицируемого риска указывает на степень, до которой можно диверсифицировать риск приобретения акций McDonald's, инвестировав часть своего портфеля в акции из списка Dow Jones Industrial.

е) Найдите уравнение линии регрессии для прогнозирования процентного изменения курса акций McDonald's на основании процентного изменения индекса Dow Jones Industrial. Определите так называемый коэффициент "бета" для акций — меру, используемую биржевыми аналитиками, — которая равняется наклону этой линии. В соответствии с моделью определения цен фиксированных активов акции с большими значениями "бета", как правило, обеспечивают более крупные ожидаемые доходы (в среднем, с течением времени), чем акции с меньшими значениями "бета".

ж) Найдите 95% доверительный интервал для коэффициента наклона.

з) Проверьте на уровне 5%, существует ли значимая связь между суточными процентными изменениями курса акций McDonald's и индекса Dow Jones.

и) Проверьте на уровне 5%, значимо ли отличается от 1 "бета" акций McDonald's ("бета", равная 1, соответствует высоко диверсифицированному портфелю).

8. Продолжим анализ биржевых данных о McDonald's и Dow Jones.

а) Найдите 95% доверительный интервал для процентного изменения курса акций McDonald's в тот день, когда индекс Dow Jones оставался неизменным.

б) Найдите 95% доверительный интервал для среднего процентного изменения курса акций McDonald's для идеализированной генеральной совокупности всех таких дней, когда индекс Dow Jones оставался неизменным.

в) Найдите 95% доверительный интервал для процентного изменения курса акций McDonald's в день, когда индекс Dow Jones поднялся на 1,5%.

г) Найдите 95% доверительный интервал для среднего процентного изменения курса акций McDonald's для идеализированной генеральной совокупности всех таких дней, когда индекс Dow Jones поднимался на 1,5%.

Таблица 11.3.5. Суточные изменения биржевых курсов за ноябрь и декабрь 1998 г. (в процентах)

McDonald's	Dow Jones	McDonald's	Dow Jones
1,33	0,84	–0,76	–0,35
0,00	–2,77	–2,04	–1,77
0,88	3,71	1,54	1,17
1,51	–0,55	0,60	0,00
0,67	1,75	–0,47	2,23
–0,86	–2,81	–0,21	–1,75
–0,38	6,15	–1,86	–3,64
–0,45	–2,37	–0,22	0,74
0,07	1,71	–1,43	–2,20
1,02	0,97	1,47	1,40
1,03	1,14	–0,37	3,60
–0,28	–3,90	0,97	2,58
0,61	0,00	0,31	4,52
0,17	2,43	0,96	3,08
1,14	0,09	0,62	–0,81
2,34	1,85	1,74	0,00
–0,78	–0,60	0,17	–0,73
0,14	0,39	0,10	–1,31
0,20	0,00	1,02	2,24
–2,32	–2,94	–0,50	1,70
0,19	0,89	–1,00	–1,92

9. В примере о зависимости объема продаж от размера территории (основанном на данных из табл. 11.2.3) оказалось, что линия наименьших квадратов, позволяющая прогнозировать объем продаж в зависимости от численности населения на соответствующей территории, определяется следующим выражением:

ожидаемый объем продаж = \$1 371 744 + \$0,23675045 (численность населения).

а) Интерпретируйте коэффициент наклона как число, имеющее простой и непосредственный экономический смысл.

б) Какая часть вариации объема продаж разных менеджеров объясняется величиной территории? Какая часть объясняется действием других факторов?

в) Насколько существенно влияние величины территории на объем продаж? Поясните свой ответ.

г) Найдите p-значение (в виде $p > 0,05$; $p < 0,05$; $p < 0,01$ или $p < 0,001$) для значимости коэффициента наклона.

10. Уравнение прогнозирования, построенное методом наименьших квадратов, имеет такой вид: прогнозируемые затраты = 35,2 + 5,3(количество изделий), причем прогнозируемые затраты измеряются в долларах. Найдите прогнозируемое значение и остаток для ситуации, когда затраты равны \$600, а количество изделий — 100.

11. Найдите значение из t-таблицы, которое будет использоваться при построении доверительного интервала для коэффициента наклона в регрессионном анализе для каждой из перечисленных ниже ситуаций.

а) Для 95% доверительного интервала при размере выборки $n = 298$.

б) Для 99% доверительного интервала при размере выборки $n = 15$.

в) Для 95% доверительного интервала при размере выборки $n = 25$.

г) Для 99,9% доверительного интервала при размере выборки $n = 100$.

12. В табл. 11.3.6 указаны вес и цена золотых монет.

а) Насколько сильна связь между весом и ценой для этих монет? Вычислите значение и дайте его словесную интерпретацию.

б) Найдите уравнение регрессии для прогноза цены на основании веса.

Таблица 11.3.6. Золотые монеты

Название	Вес, тройские унции	Цена, дол.
Кленовый лист	1	400,75
Мексиканская	1,2	482,00
Австралийская	0,9802	382,00
Американский орел	1	400,75
Американский орел	0,5	210,75
Американский орел	0,25	114,00
Американский орел	0,1	53,00

в) Интерпретируйте коэффициент наклона как имеющий реальный смысл показатель цены.

г) Насколько примерно различаются прогнозируемые и фактические цены (в долларах)?

д) Найдите 95% доверительный интервал для коэффициента наклона.

е) Отличается ли коэффициент наклона значимо от 0? Почему вы так считаете?

13. Какова площадь пригодных для застройки участков на территории Сиэтла? В табл. 11.3.7 указаны количество существующих домов и возможности застройки свободной земли в ряде районов Сиэтла.

а) Насколько сильна связь между количеством существующих домов и возможностью застройки свободной земли? Укажите конкретное число и дайте его словесную интерпретацию.

б) Найдите уравнение регрессии, позволяющее прогнозировать возможности застройки свободных земель на основании количества существующих домов.

Таблица 11.3.7. Площадь под застройку

Район	Существующие дома	Возможность застройки свободной земли
Bear Creek	6 800	11 814
East Sammamish	10 900	5 800
Eastside	600	153
Federal Way	11 200	5 340
Green R. Valley	1 050	186
Highline	33 600	8 265
Newcastle	16 700	7 421
Northshore	24 500	6 474

Данные из *Seattle Post-Intelligencer*, 1991, December 4, p. A8. Использован источник King County.

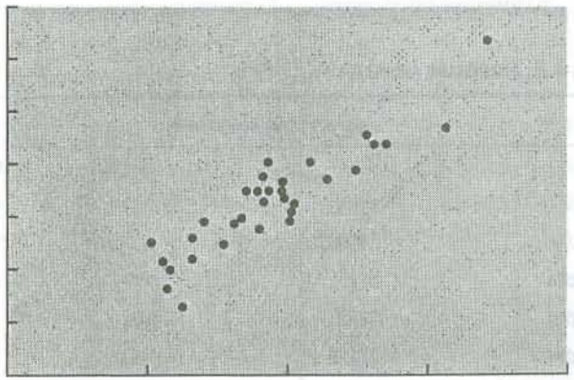


Рис. 11.3.1

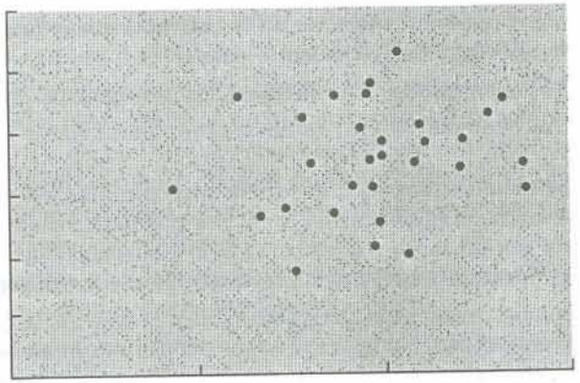


Рис. 11.3.2

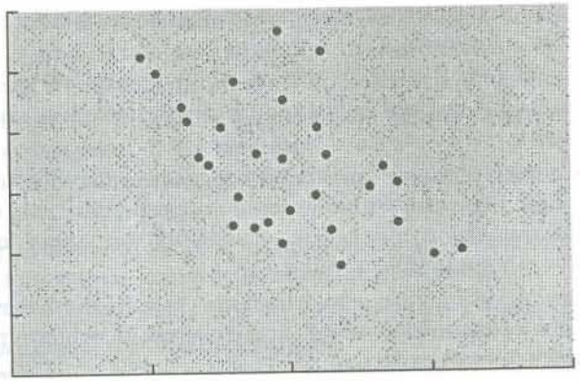


Рис. 11.3.3

в) Найдите значение остатка для района Highline, прогнозируя возможности застройки свободной земли на основании количества существующих домов.

г) Найдите 95% доверительный интервал для коэффициента наклона.

д) Наблюдается ли значимая взаимосвязь между возможностями застройки и количеством существующих домов? Поясните свой ответ.

14. Для каждой из диаграмм рассеяния, показанных на рис. 11.3.1–11.3.3, укажите, к какому из перечисленных ниже значений ближе всего значение корреляции: 0,9; 0,5; 0,0; −0,5 или −0,9.

15. Рассмотрим розничную цену неэтилированного бензина, продающегося на автозаправочных станциях в разных местах и в разное время (соответствующие данные приведены в табл. 11.3.8).

а) Насколько сильна связь между ценами в ноябре и в феврале? Укажите конкретное число и дайте его словесную интерпретацию.

б) Найдите уравнение регрессии, позволяющее прогнозировать более поздние цены на основании цен, действовавших ранее.

в) Найдите значение остатка для Линвуда (прогнозируя более поздние цены на основании цен, действовавших ранее).

г) Найдите 95% доверительный интервал для коэффициента наклона.

д) Значимо ли отличается этот коэффициент наклона от 0? Почему вы так считаете?

16. Время от времени в средствах массовой информации обсуждаются высокие должностные оклады президентов и других руководителей различных благотворительных организаций. В табл. 11.3.9 приведена информация о благотворительной организации United Way в 10 крупнейших городах страны.

а) Какой процент вариации должностных окладов президентов объясняется тем, что некоторым из них удается собрать больше пожертвований в расчете на душу населения, чем другим? Укажите конкретное число и дайте общепринятое статистическое название этого понятия.

б) Найдите уравнение регрессии, позволяющее прогнозировать величину должностного оклада на основании собранной суммы пожертвований в расчете на душу населения.

в) Найдите значение остатка для Сиэтла, прогнозируя величину должностного оклада на основании собранной суммы пожертвований в расчете на душу населения.

г) Найдите показатель, который характеризовал бы типичную ошибку, допускаемую при использовании уравнения регрессии для прогнозирования величины должностного оклада на основании собранной суммы пожертвований в расчете на душу населения.

д) Наблюдается ли значимая взаимосвязь между должностными окладами президентов и собранной суммой пожертвований в расчете на душу населения? Поясните свой ответ.

17. В табл. 11.3.10 приведены данные о величине списка почтовой рассылки (в тысячах фамилий) и объеме продаж (в тысячах долларов) по группе каталогов.

а) Насколько сильна связь между этими двумя переменными? Найдите соответствующий показатель и интерпретируйте его.

б) Найдите уравнение регрессии, позволяющее прогнозировать объем продаж на основании величины списка почтовой рассылки.

в) На какой уровень продаж можно рассчитывать в случае каталога, располагающего 5000 подписчиками?

Таблица 11.3.8. Цены на бензин

Район	Цена на 11/30/90, дол.	Цена на 02/26/91, дол.
Сиэтл	136,9	114,0
Бельвью	138,7	113,6
Эверетт	138,7	114,5
Линвуд	137,1	110,3
Рентон	137,9	112,7

г) Какой процент вариации размера списка можно объяснить тем, что некоторые из этих списков обеспечивают больший объем продажи, чем другие?

д) Наблюдается ли значимая взаимосвязь между величиной списка почтовой рассылки и объемом продаж? Поясните свой ответ.

18. В табл. 11.3.11 фонды краткосрочных облигаций сравниваются по таким показателям, как средний срок погашения облигаций (измеряется в количестве лет до наступления срока погашения облигаций фонда) и дивиденды (в процентном выражении).

а) Определите корреляцию между сроками погашения облигаций и дивидендами и интерпретируйте эту корреляцию.

б) Постройте методом наименьших квадратов уравнение регрессии, позволяющее прогнозировать дивиденды на основании срока погашения облигаций.

в) На какие дивиденды можно рассчитывать при покупке облигаций фонда, срок погашения которых на данный момент составляет ровно один год?

г) Найдите стандартную ошибку прогнозирования (для прогнозирования "дивидендов" при заданном сроке погашения облигаций) и дайте ее содержательную интерпретацию.

д) Наблюдается ли значимая взаимосвязь между сроком погашения облигаций и дивидендами? Поясните свой ответ.

19. В табл. 11.3.12 представлены данные о суточном объеме производства и количестве занятых работников для некоторой совокупности дней.

а) Найдите уравнение регрессии, позволяющее прогнозировать объем производства, исходя из количества занятых работников.

Таблица 11.3.9. Благотворительные организации

Город	Должностной оклад президента, дол.	Собранная сумма пожертвований (в расчете на душу населения), дол.	Город	Должностной оклад президента, дол.	Собранная сумма пожертвований (в расчете на душу населения), дол.
Атланта	161 396	17,35	Хьюстон	146 641	15,89
Чикаго	189 808	15,81	Канзас-Сити	126 002	23,87
Кливленд	171 798	31,49	Лос-Анджелес	155 192	9,32
Денвер	108 364	51,51	Миннеаполис	169 999	29,84
Детройт	201 490	16,74	Сиэтл	143 025	24,19

Таблица 11.3.10. Списки почтовой рассылки

Величина списка	Объем продаж	Величина списка	Объем продаж
168	5 178	249	7 325
21	2 370	43	2 449
94	3 591	589	15 708
39	2 056	41	2 469

б) Какой будет оценка объема производства, обеспечиваемого одним дополнительным работником?

в) Изобразите диаграмму рассеяния и линию регрессии для рассматриваемой совокупности данных.

г) Найдите ожидаемый объем производства и остаток для первой пары значений. Интерпретируйте оба полученных значения с экономической точки зрения.

д) Найдите стандартную ошибку коэффициента наклона. Что означает это число?

е) Найдите 95% доверительный интервал для ожидаемой предельной ценности дополнительного работника. (Так на языке экономистов называется наклон.)

ж) Проверьте на уровне 5%, существует ли значимая взаимосвязь между объемом производства и количеством работников.

20. Известны коэффициент корреляции, $r = -0{,}603$, и построенное методом наименьших квадратов уравнение прогнозирования — $Y = 38{,}2 - 5{,}3X$. Найдите прогнозируемое значение Y при $X = 15$.

Таблица 11.3.11. Фонды краткосрочных облигаций

Фонд	Срок погашения облигаций	Дивиденды, %
Strong Short-Term Bond Fund	1,11	7,43
DFA One-Year Fixed-Income Portfolio	0,76	5,54
Scudder Target Government Zero-Coupon 1990	2,3	5,01
IAI Reserve Fund	0,4	4,96
Scudder Target Fund General 1990	1,9	4,86
Vanguard Fixed-Income Short-Term Bond Portfolio	2,3	4,86
Criterion Limited-Term Institutional Trust	1,3	4,8
Franklin Series Trust Short-Int. U.S. Govt.	2	4,64
Benham Target Maturities Trust Series 1990	2,3	4,62
Delaware Treasury Reserves Investors Series	2,84	4,35

Таблица 11.3.12. Суточный объем производства

Количество работников	Объем производства	Количество работников	Объем производства
7	483	9	594
6	489	9	575
7	486	6	464
8	562	9	647
8	568	8	595
9	559	6	499

21. Известны коэффициент корреляции, $r = 0{,}307$, и построенное методом наименьших квадратов уравнение прогнозирования — $Y = 55{,}6 + 18{,}2X$. Найдите прогнозируемое значение Y при $X = \$25$.
22. В один из дней на вашем заводе для производства 132 изделий было израсходовано электроэнергии на сумму \$385. В другой день для производства 183 изделий было израсходовано электроэнергии на сумму \$506. На третий день для производства 105 изделий было израсходовано электроэнергии на сумму \$261. Дайте оценку, сколько, по вашему мнению, будет израсходовано электроэнергии для производства 150 изделий?
23. Какой из перечисленных ниже коэффициентов корреляции соответствует умеренно сильной взаимосвязи, при которой большим значениям Y соответствуют большие значения X: $r = 1$; $r = 0{,}73$; $r = 0{,}04$; $r = -0{,}83$ или $r = -0{,}99$?
24. В понедельник ваше предприятие выпустило 7 изделий, которые обошлись вам в \$18. Во вторник вы выпустили 8 изделий стоимостью \$17, в среду — 18 изделий стоимостью \$32, в четверг — 3 изделия стоимостью \$16. Воспользовавшись моделью линейной регрессии, учитывающей фиксированные и переменные затраты, оцените, во сколько вам обойдется выпуск 10 изделий в пятницу.
25. На выходных днях (в конце недели) вы снизили цены на 5%, и объем продаж в вашем магазине составил \$58 000. Во время следующих выходных дней вы снизили цены на 15%, и объем продаж в вашем магазине достиг \$92 000. Затем во время следующих выходных дней вы снизили цены на 17,5%, и объем продаж достиг \$95 000. Основываясь на этой информации, оцените ожидаемый объем продаж во время будущих выходных дней, после того как вы снизите цены на 10%.
26. Определите структуру точек диаграммы рассеяния на рис. 11.3.4.
27. Определите структуру точек диаграммы рассеяния на рис. 11.3.5.
28. Рассмотрим доходы на одну акцию и курс акций, регистрируемый в конце рабочего дня для некоторых фирм со значительной рыночной капитализацией, работающих в области биотехнологий. Учитывая важность, придаваемую многими аналитиками величине доходов на одну акцию, можно предположить наличие сильной корреляции между величиной доходов на одну акцию и курсом акций. Разумеется, в такой сравнительно новой и бурно развивающейся отрасли, как биотехнология, делать далеко идущие выводы было бы пока опрометчиво, поскольку биржевой курс может в значительной мере зависеть не столько от фактически достигнутых доходов, сколько от ожиданий будущих доходов (имеющих случайный характер). Давайте посмотрим, как это происходит. Соответствующие данные представлены в табл. 11.3.13.

а) Изобразите диаграмму рассеяния для зависимости курса акций от величины доходов на одну акцию.

б) Найдите коэффициент корреляции. Какого рода взаимосвязь — положительная или отрицательная — наблюдается между доходами и курсом акций?

в) Проверьте на уровне 5%, существует ли для этих фирм значимая взаимосвязь между величиной доходов на одну акцию и курсом акций.

г) Проверьте на уровне 10%, существует ли для этих фирм значимая взаимосвязь между величиной доходов на одну акцию и курсом акций.

д) Кратко опишите результаты анализа.

е) Вы являетесь главой фирмы, занимающейся биотехнологиями, которая собирается вскоре приступить к свободной продаже своих акций. Ваши доходы на одну акцию составляют $0,05. Основываясь исключительно на проведенном к настоящему времени регрессионном анализе, укажите, на какой курс акций вы можете рассчитывать.

ж) Вычислите двусторонний 95% доверительный интервал для этого прогнозируемого курса акций.

з) Вычислите двусторонний 95% доверительный интервал для среднего ожидаемого курса акций для генеральной совокупности всех таких фирм с доходом $0,05 на одну акцию.

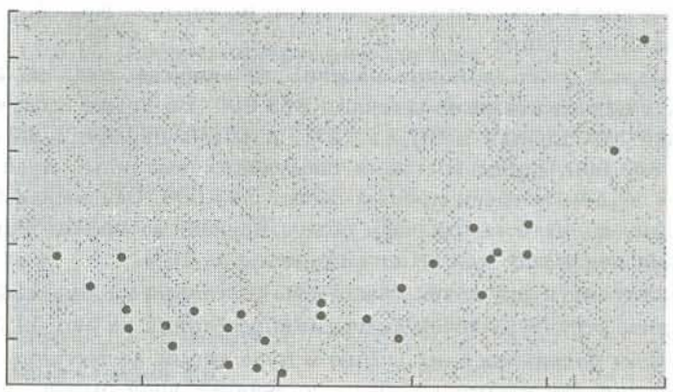


Рис. 11.3.4

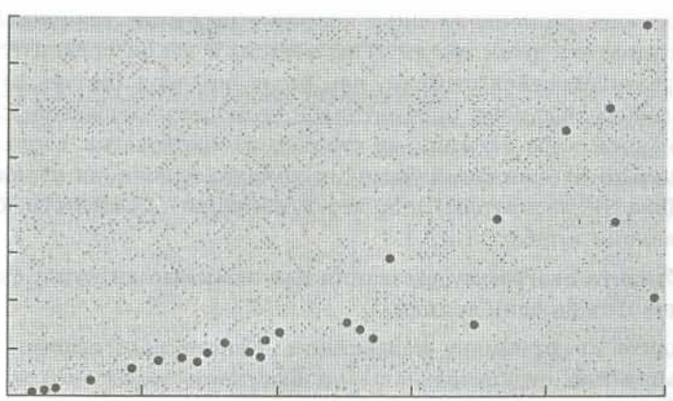


Рис. 11.3.5

Таблица 11.3.13. Акции фирм, работающих в области биотехнологий

	Доходы на одну акцию, дол.	Цена на 01/27/89, дол.
Genentech	0,24	17,88
Alza	0,50	24,75
Amgen	0,09	37,00
Cetus	–0,22	11,38
Genetics Institute	–0,81	18,75
Biogen	–0,21	9,38
Centocor	0,21	17,00
Chiron	–0,97	15,00
Xoma	–1,00	15,00
Nova Pharmaceutical	–0,32	5,38
Immunex	0,02	11,75
Collagen	0,12	11,38
California Biotech	–0,87	5,25
Calgene	–0,66	6,38
DNA Plant Technology	–0,16	4,63
Repligen	–0,57	7,25
Imreg	–0,36	4,50
Celgene	–0,90	8,75
Cytogen	–1,10	3,63
Damon Biotech	–0,27	1,75

Данные взяты из A.K Naj "Clouds Gather over the Biotech Industry; Firms Stumble on Regulation, Patent Problems", *The Wall Street Journal*, 1989, January 30, p. B1. Источник данных: IDD Information Services/Tradeline.

29. Рассмотрим тираж издания и тариф на размещение рекламы (цена строки одноразового рекламного объявления). Соответствующие данные для нескольких крупных газет приведены в табл. 11.3.14.

а) Постройте диаграмму рассеяния для зависимости тарифа на размещение рекламы от тиража газеты.

б) Найдите и интерпретируйте корреляцию тарифа на размещение рекламы и тиража газеты. Можно ли объяснить эту корреляцию с экономической точки зрения?

в) Найдите уравнение регрессии, позволяющее прогнозировать тариф на размещение рекламы на основании тиража газеты.

г) Проверьте, является ли связь между тарифом на размещение рекламы и тиражом газеты статистически значимой.

д) Найдите прогнозируемое значение и остаток для *New York Times*. Интерпретируйте эти значения. Ответьте, в частности, на вопрос, является ли

Таблица 11.3.14. Цена строки рекламы в крупнейших газетах

	Тираж, экз.	Цена строки одноразовой рекламы, дол.
The Wall Street Journal	2 081 995	37,65
New York Daily News	1 374 858	18,48
USA Today	1 284 613	14,50
Los Angeles Times	1 057 536	14,61
New York Times	970 051	16,47
New York Post	963 069	16,07
Philadelphia Inquirer/News	828 236	13,82
Chicago Tribune	779 259	13,05
Washington Post	768 288	13,78
San Francisco Chronicle/Examiner	691 771	12,25
Chicago Sun-Times	663 693	10,53
Detroit News	657 015	14,18
Detroit Free Press	645 623	12,83
Long Island Newsday	533 384	7,81
Kansas City Times/Star	528 777	5,17
Miami Herald/News	514 702	11,08
Cleveland Plain Dealer	492 002	6,58
Milwaukee Journal	486 426	8,77
Houston Chronicle	443 592	6,03
Baltimore Sun	349 182	6,77

Данные взяты из Bovée C. L. and Arens W. F. *Contemporary Advertising*, 2nd ed. (Homewood, Ill.: Richard D. Irwin, 1986), p. 413.

тариф на размещение одной строки рекламного объявления выше или ниже той величины, которую можно было бы ожидать для газеты с таким тиражом.

30. В предыдущей задаче на диаграмме рассеяния виден один возможный выброс. Давайте выясним, принадлежит ли это значение той же генеральной совокупности, что и остальные, рассматривая его как новое наблюдение.[26]

а) Удалите *The Wall Street Journal* из рассматриваемой совокупности данных и найдите уравнение регрессии, позволяющее прогнозировать тариф на размещение рекламы на основании тиража для всех остальных газет.

[26] Более строгий анализ позволяет учесть тот факт, что газету *The Wall Street Journal* действительно нельзя рассматривать как результат случайного выбора из некоторой идеализированной генеральной совокупности, поскольку она была выбрана, в частности, именно с учетом ее большого отличия от других газет. Подробный анализ проблем, связанных с резко отклоняющимися значениями, и методы их решения приведены в книге Barnett V. and Lewis T. *Outliers in Statistical Data* (New York: Wiley, 1978).

б) Вычислите двусторонний 95% доверительный интервал для нового наблюдения, когда X_0 является размером тиража *The Wall Street Journal*.

в) Проверьте, является ли *The Wall Street Journal* выбросом, выяснив, попадает ли тариф на размещение рекламы в этой газете в построенный доверительный интервал.

31. Тариф за миллион для рекламы в газетах определяется как тариф на размещение одной строки рекламы, деленный на тираж (в миллионах экземпляров). Таким образом, этот тариф представляет собой затраты на одну строку рекламного объявления на каждый миллион экземпляров тиража. Эта поправка должна учитывать некоторые различия в тарифах на размещение рекламы, обусловленные тиражом. Таким образом, одно из объяснений тарифа на размещение одной строки рекламы заключается в том, что он пропорционален тиражу. Если же он действительно пропорционален тиражу, тогда в тарифе за миллион не остается ничего, что должно объясняться тиражом. С другой стороны, если в большом тираже есть какое-то дополнительное преимущество (или недостаток), то величина тиража должна помочь объяснить вариацию тарифов за миллион. Давайте попытаемся с помощью регрессионного анализа выяснить, остается ли в тарифе за миллион что-то такое, что объясняется тиражом. Воспользуемся данными, приведенными в табл. 11.3.15.

а) Постройте диаграмму рассеяния тарифа за миллион в зависимости от тиража.

б) Найдите и интерпретируйте корреляцию между тиражом и тарифом за миллион.

в) Какой процент вариации тарифа за миллион объясняется тиражом?

г) Проверьте, существует ли значимая взаимосвязь между тиражом и тарифом за миллион.

д) Кратко поясните и интерпретируйте в письменном виде полученные вами результаты.

32. Ваша фирма, выпускающая ряд пластмассовых деталей для автомобилей, не может добиться нужного уровня качества своей продукции (слишком большой процент брака). Один из ваших инженеров полагает, что причиной этого является недостаточно тщательный контроль температуры соответствующих технологических процессов. Другому инженеру кажется, что все дело в слишком частых остановках сборочной линии, которые происходят по не связанным между собой причинам. Вы решили проанализировать проблему и собрать необходимые для этого данные. В табл. 11.3.16 содержатся данные о проценте брака за несколько последних дней, о стандартном отклонении температуры, измерявшейся каждый час на протяжении этих дней (эти данные служат мерой контроля температуры), и о количестве остановок сборочной линии за каждый из этих дней.

а) Найдите корреляцию между процентом брака и изменчивостью температуры.

б) Найдите корреляцию между процентом брака и остановками сборочной линии.

Таблица 11.3.15. Тарифы на размещение одной строки рекламных объявлений в газетах с поправкой на тираж

	Тираж, экз.	Тариф за миллион, дол.
The Wall Street Journal	2 081 995	18,08
New York Daily News	1 374 858	13,43
USA Today	1 284 613	11,29
Los Angeles Times	1 057 536	13,81
New York Times	970 051	16,98
New York Post	963 069	16,69
Philadelphia Inquirer/News	828 236	16,69
Chicago Tribune	779 259	16,75
Washington Post	768 288	17,94
San Francisco Chronicle/Examiner	691 771	17,71
Chicago Sun-Times	663 693	15,87
Detroit News	657 015	21,58
Detroit Free Press	645 623	19,87
Long Island Newsday	533 384	14,64
Kansas City Times/Star	528 777	9,78
Miami Herald/News	514 702	21,53
Cleveland Plain Dealer	492 002	13,37
Milwaukee Journal	486 426	18,03
Houston Chronicle	443 592	13,59
Baltimore Sun	349 182	19,39

Данные взяты из Bovée C. L. and Arens W. F. *Contemporary Advertising*, 2nd ed. (Homewood, Ill.: Richard D. Irwin, 1986), p. 413.

в) Какая из возможных причин — изменчивость температуры или остановки сборочной линии — сказывается в большей степени на вариации процента брака в разные дни? Поясните свой ответ.

г) Оцените статистическую значимость каждого из найденных вами коэффициентов корреляции.

д) Изобразите диаграмму рассеяния для процента брака в зависимости от количества остановок сборочной линии. Кратко интерпретируйте в письменном виде полученную вами диаграмму и коэффициент корреляции.

е) Изобразите диаграмму рассеяния значений процента брака в зависимости от изменчивости температуры. Кратко интерпретируйте в письменном виде полученную вами диаграмму и коэффициент корреляции.

ж) Кратко резюмируйте в письменном виде свои выводы из полученных результатов и предложения, касающиеся повышения качества выпускаемых изделий.

Таблица 11.3.16. Брак продукции и его возможные причины

Процент брака	Изменчивость температуры	Остановки сборочной линии	Процент брака	Изменчивость температуры	Остановки сборочной линии
0,1	11,94	5	0,0	10,10	2
0,1	9,33	4	5,2	13,08	2
8,4	21,89	0	4,9	17,19	0
0,0	8,32	1	0,1	10,76	1
4,5	14,55	0	6,8	13,73	3
2,6	12,08	8	4,8	12,42	2
3,2	12,16	0	0,0	12,83	2
0,0	12,56	2	0,9	5,78	5

Упражнения с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении А.

1. Рассмотрим годовую заработную плату как переменную Y, а стаж работы — как переменную X.

   а) Постройте диаграмму рассеяния и опишите взаимосвязь

   б) Найдите коэффициент корреляции. О чем свидетельствует его значение? Соответствует ли он диаграмме рассеяния?

   в) Найдите методом наименьших квадратов линию регрессии для прогноза Y на основании X и начертите ее на диаграмме рассеяния.

   г) Найдите стандартное отклонение оценки. О чем свидетельствует этот показатель?

   д) Найдите стандартную ошибку коэффициента наклона.

   е) Найдите 95% доверительный интервал для коэффициента наклона.

   ж) Проверьте на уровне 5% значимость отличия от 0 коэффициента наклона. Интерпретируйте полученный результат.

   з) Проверьте на уровне 1% значимость отличия от 0 коэффициента наклона.

   и) Проверьте на уровне 5% значимость отличия от 0 коэффициента корреляции.

2. а) Какая доля вариации заработной платы объясняется тем, что одни служащие имеют больший стаж работы, чем другие? 

   б) Какую заработную плату можно ожидать для служащего со стажем работы 8 лет?

   в) Найдите 95% доверительный интервал для заработной платы нового служащего со стажем работы 8 лет, взятого из той же генеральной совокупности, из которой были извлечены исследуемые нами данные.

г) Найдите 95% доверительный интервал для средней заработной платы всех тех служащих в генеральной совокупности, которые имеют стаж работы 8 лет.

3. а) Какую заработную плату можно ожидать для служащего с трехлетним стажем работы?

   б) Найдите 95% доверительный интервал для заработной платы нового служащего с трехлетним стажем работы, взятого из той же генеральной совокупности, из которой были извлечены исследуемые нами данные.

   в) Найдите 95% доверительный интервал для средней заработной платы всех тех служащих в генеральной совокупности, которые имеют стаж работы 3 года.

4. а) На какую заработную плату может рассчитывать служащий с нулевым стажем работы (т.е. без стажа работы)?

   б) Найдите 95% доверительный интервал для заработной платы нового служащего без стажа работы, взятого из той же генеральной совокупности, из которой были извлечены исследуемые нами данные.

   в) Найдите 95% доверительный интервал для средней заработной платы всех тех служащих в генеральной совокупности, которые имеют нулевой стаж работы.

5. Будем рассматривать годовую заработную плату как переменную Y, а возраст — как переменную X.

   а) Постройте диаграмму рассеяния и опишите взаимосвязь.

   б) Найдите коэффициент корреляции. О чем свидетельствует значение этого коэффициента? Соответствует ли он диаграмме рассеяния?

   в) Найдите методом наименьших квадратов линию регрессии для прогнозирования Y на основании X и начертите ее на вашей диаграмме рассеяния.

   г) Найдите стандартную ошибку оценки. О чем свидетельствует этот показатель?

   д) Найдите стандартную ошибку коэффициента наклона.

   е) Найдите 95% доверительный интервал для коэффициента наклона.

   ж) Проверьте на уровне 5% значимость отличия коэффициента наклона от 0. Интерпретируйте полученный результат.

   з) Проверьте на уровне 1% значимость отличия коэффициента наклона от 0.

6. а) Какая доля вариации заработной платы объясняется тем, что у одних служащих возраст больше, чем у других?

   б) Какую заработную плату можно ожидать для служащего в возрасте 42 лет?

   в) Найдите 95% доверительный интервал для заработной платы нового служащего в возрасте 42 лет, взятого из той же генеральной совокупности, из которой были извлечены исследуемые нами данные.

   г) Найдите 95% доверительный интервал для средней заработной платы всех тех служащих в рассматриваемой генеральной совокупности, возраст которых составляет 42 года.

7. а) Какую заработную плату можно ожидать для служащего в возрасте 50 лет?

   б) Найдите 95% доверительный интервал для заработной платы нового служащего в возрасте 50 лет, взятого из той же генеральной совокупности, из которой были извлечены исследуемые нами данные.

   в) Найдите 95% доверительный интервал для средней заработной платы всех тех служащих в рассматриваемой генеральной совокупности, возраст которых составляет 50 лет.

8. Будем рассматривать стаж работы как переменную Y, а возраст — как переменную X.

   а) Постройте диаграмму рассеяния и опишите взаимосвязь.

   б) Найдите коэффициент корреляции. О чем свидетельствует значение этго коэффициента? Соответствует ли он вашей диаграмме рассеяния?

   в) Найдите методом наименьших квадратов линию регрессии для прогнозирования Y на основании X и начертите ее на вашей диаграмме рассеяния.

   г) Найдите стандартную ошибку оценки. О чем свидетельствует этот показатель?

   д) Найдите стандартную ошибку коэффициента наклона.

   е) Найдите 95% доверительный интервал для коэффициента наклона.

   ж) Проверьте на уровне 5% значимость отличия коэффициента наклона от 0. Интерпретируйте полученный результат.

   з) Проверьте на уровне 1% значимость отличия коэффициента наклона от 0.

9. а) Какая доля вариации стажа работы объясняется тем, что у одних служащих возраст больше, чем у других?

   б) Можно ожидать наличие стажа работы у служащего в возрасте 42 лет?

   в) Найдите 95% доверительный интервал для стажа работы нового служащего в возрасте 42 лет, взятого из той же генеральной совокупности, из которой были получены исследуемые нами данные.

   г) Найдите 95% доверительный интервал для среднего стажа работы всех тех служащих в генеральной совокупности, возраст которых составляет 42 года.

10. а) Наличие какого стажа работы можно ожидать у служащего в возрасте 50 лет?

    б) Найдите 95% доверительный интервал для стажа работы нового служащего в возрасте 50 лет, взятого из той же генеральной совокупности, из которой были получены исследуемые нами данные.

    в) Найдите 95% доверительный интервал для среднего стажа работы всех тех служащих в генеральной совокупности, возраст которых составляет 50 лет.

Проекты



Используя Internet, газеты или журналы, подберите двумерную совокупность данных с размером выборки $n = 15$ или больше, касающуюся вашей работы или бизнеса.

1. Выберите зависимую переменную (Y) и независимую переменную (X) и кратко обоснуйте свой выбор.

2. Изобразите соответствующую диаграмму рассеяния и сделайте комментарии относительно взаимосвязи.
3. Вычислите коэффициент корреляции и кратко интерпретируйте полученное значение.
4. Возведите коэффициент корреляции в квадрат и кратко интерпретируйте полученное значение.
5. Найдите методом наименьших квадратов уравнение регрессии и начертите соответствующую линию на диаграмме рассеяния.
6. Вычислите прогнозируемые значения Y и остатки для двух объектов вашей совокупности данных. Сделайте комментарий.
7. Найдите доверительный интервал для коэффициента наклона.
8. Проверьте, можно ли что-нибудь объяснить на основании полученного вами уравнения регрессии.
9. Выберите какое-либо значение X. Найдите ожидаемое значение Y для этого X. Найдите доверительный интервал для значения Y у объекта с этим значением X. Найдите доверительный интервал для среднего значения Y у генеральной совокупности объектов с этим значением X. Обобщите и проинтерпретируйте полученные результаты.
10. Сделайте выводы из результатов применения корреляционного и регрессионного анализа к этой совокупности данных. Что нового вы узнали об исследуемой совокупности данных?

Ситуация для анализа

Еще один этап производства: нужен ли он?

Специалисты из научно-исследовательской лаборатории предлагают вам (и руководству компании в целом) добавить в производственный процесс еще один этап. Они увлечены этой идеей, однако вас одолевают сомнения, поскольку всем известно, что один из ее авторов является приятелем главы компании, работающей в области биотехнологий и занимающейся производством реагента, который предполагается использовать на дополнительном этапе производства. Но если внедрение нового этапа даст такие результаты, как ожидается, это должно помочь вашей компании существенно сократить производственные расходы. Проблема, однако, заключается в том, что только что полученные результаты испытаний не вселяют в специалистов вашей компании чересчур большого оптимизма. В связи с этим предполагается провести совещание технических специалистов и руководства компании. Желая получить перед этим совещанием максимально объективную информацию, вы решаете самостоятельно проанализировать имеющиеся данные.

Ваша компания рассчитывает получить со стороны Комитета по продуктам питания и лекарственным препаратам (Food and Drug Administration — FDA) разрешение на продажу нового медицинского диагностического теста, основанного на технологии моноклональных антител, а вы входите в группу специалистов, ответственных за производство. Естественно, эта группа занимается исследованием способов повышения объемов производства и сокращения расходов.

Суть рассматриваемого предложения сводится к добавлению в технологический процесс еще одной реакции, обеспечивающей промежуточную очистку продукта. Такой подход следует признать рациональным, поскольку ресурсы концентрируются на последних стадиях производства продукта. Однако он порождает проблему, связанную с любым дополнительным этапом производственного процесса: еще один вид обработки, еще одно вмешательство, еще один источник возможных сложностей и проблем. Что касается рассматриваемого нами случая, то высказывалось следующее предположение: в то время как небольшие количества упоминавшегося нами реагента могут действительно принести пользу, попытки выполнить слишком глубокую очистку на самом деле приведут лишь к снижению объемов производства и повышению производственных расходов.

Суть предлагаемой проверки заключается в проведении ряда производственных циклов, в каждом из которых используются разные количества вещества-очистителя, причем в одном из производственных циклов этап очистки исключался полностью (т.е. количество вещества-очистителя равно 0). Последовательность испытаний должна носить случайный характер, чтобы какие-либо временные тенденции ошибочно не воспринимались так, словно они вызваны процедурой дополнительной очистки. Ниже приведены соответствующие данные, а также результаты регрессионного анализа.

Количество вещества-очистителя	Наблюдаемый объем производства	Количество вещества-очистителя	Наблюдаемый объем производства
0	13,39	6	37,07
1	11,86	7	51,07
2	27,93	8	51,69
3	35,83	9	31,37
4	28,52	10	21,26
5	41,21		

Итоговый отчет

Статистические характеристики регрессии	
Множественный R	0,516
R-квадрат	0,266
R-квадрат с поправкой	0,184
Стандартная ошибка	12,024
Наблюдения	11

ANOVA

	df	SS	MS	F	Значимость F
Регрессия	1	471,156	471,156	3,259	0,105
Остаток	9	1301,294	144,588		
Итого	10	1772,450			

Окончание таблицы

	Коэффициенты	Стандартная ошибка	t	P	Нижний 95%	Верхний 95%
Сдвиг	21,578	6,783	3,181	0,011	6,234	36,922
Вещество-очиститель	2,070	1,146	1,805	0,105	–0,524	4,663

Вопросы для обсуждения

1. Можно ли сказать, исходя из проведенного регрессионного анализа, что объем вещества-очистителя оказывает существенное влияние на объем производства? Можно ли, основываясь на вашем ответе на данный вопрос, рекомендовать включение этапа очистки в производственный процесс?
2. Что бы вы порекомендовали? Есть ли какие-то соображения, способные изменить вашу точку зрения?

ГЛАВА **12**

Множественная регрессия: прогнозирование одного фактора на основе нескольких других

Окружающий нас мир многомерен. В подавляющем большинстве реальных экономических задач приходится рассматривать данные более чем об одном или двух факторах. Однако это не является неразрешимой проблемой: следующий шаг, *множественная регрессия*, представляет собой относительно несложную процедуру, которая позволяет вам расширить свои возможности за пределы простейших случаев одно- и двумерных данных. Более того, с соответствующими базовыми идеями вы уже знакомы: понятия среднего значения, изменчивости, корреляции, прогнозирования, доверительных интервалов и проверки гипотез изложены в предыдущих главах.



Прогнозирование единственной переменной Y на основании двух или нескольких переменных X называется **множественной регрессией.** Прогнозирование единственной переменной Y на основании единственной переменной X называется *простой регрессией*; о простой регрессии речь шла в предыдущей главе. Пользуясь множественной регрессией, мы преследуем, по сути, те же цели, что и в случае простой регрессии. Ниже приведен краткий обзор этих целей, сопровождаемый простыми примерами.

Первое. *Описание и понимание взаимосвязи.*

а) Рассмотрим взаимосвязь между заработной платой (Y) и рядом базовых характеристик служащих, таких как пол (X_1 представлен двумя значениями, 0 и 1 обозначают со-

ответственно мужчин и женщин), стаж работы (X_2) и образование (X_3). Описание и понимание того, как эти X-факторы влияют на Y, позволяет, например, выстраивать систему доказательств в судебных процессах, касающихся дискриминации по признаку пола. Коэффициент регрессии по признаку пола является оценкой величины разницы заработной платы между мужчинами и женщинами *с учетом поправки* на возраст и стаж работы. Даже если вашу фирму пока еще не обвиняют в дискриминации работников по признаку пола, все равно полезно было бы выполнить множественный регрессионный анализ, чтобы незначительные (пока еще!) проблемы не переросли в крупные, решать которые будет значительно сложнее.

б) Если ваша фирма участвует в конкурсе на реализацию тех или иных проектов, тогда — для тех проектов, конкурс на которые вам удалось выиграть — вы располагаете данными, касающимися фактических затрат (Y), оценки прямых трудозатрат (X_1), оценки затрат на материалы (X_2) и затрат на управленческие функции (X_3). Допустим, что предложение цены, с которым вы выходите на конкурс, кажется вам неоправданно низким. Определив взаимосвязь между фактическими затратами и оценками, сделанными ранее, на этапе переговоров о заключении контрактов, вы сможете выяснить, какие из оценок вы систематически занижаете или, наоборот, завышаете (с точки зрения их вклада в фактические затраты).

Второе. *Прогнозирование (предсказание) нового наблюдения.*

а) Глубокое понимание структуры затрат в вашей фирме может быть полезно во многих отношениях. Например, у вас может сложиться более правильное представление о том, какие дополнительные расходы следует запланировать на сезон повышенного спроса на продукцию вашей фирмы (в частности, можно учесть дополнительные затраты, связанные с выполнением сверхурочных работ). Если ваш бизнес претерпевает определенные изменения, вы должны уметь прогнозировать влияние этих изменений на структуру затрат. Лучше разбираться в структуре затрат своей фирмы вам поможет множественная регрессия затрат (Y) на каждый из потенциально значимых (на ваш взгляд) факторов, таких как количество выпускаемых изделий (X_1), количество работников (X_2) и объем сверхурочных работ (X_3). Результаты анализа, подобного этому, помогут вам принимать гораздо более продуманные решения, чем простое решение "посадить людей на сверхурочные работы на недельку-другую". Такой анализ поможет вам выявить скрытые расходы, которые обнаруживают тенденцию к возрастанию с ростом объемов сверхурочных работ, и делать более точные прогнозы фактических затрат, основанные на имеющейся у вас информации.

б) Ежемесячные объемы продаж в вашей фирме (временной ряд) могут объясняться сезонными колебаниями спроса. Один из способов анализа и прогнозирования объемов продаж заключается в использовании множественной регрессии, позволяющей объяснять объемы продаж (Y) на основании некоторого тренда (например, $X_1 = 1, 2, 3, \ldots$, указывающего месяцы от начала регистрации объемов продаж) и переменной для каждого месяца (например, X_2 равняется 1 для января и 0 в противном случае, X_3 представляет февраль, и т.д.). Множественную регрессию можно использовать для прогнозирования объемов продаж на несколько месяцев вперед, а также для уяснения долгосрочных тенденций и понимания, в какие месяцы объемы продаж, как правило, оказываются больше, чем в другие.

Третье. *Регулирование и управление процессом.*

а) На вход технологической цепочки, используемой на целлюлозно-бумажном комбинате, поступает целлюлозная масса, а на выходе получается готовая к употреблению бумага. Как управлять столь сложным комплексом оборудования? Одного лишь внимательного изучения технической документации явно недостаточно — чтобы научиться правильно регулировать технологический процесс (с точки зрения минимизации расхода электроэнергии), нужны многие годы практического опыта. Если этот опыт выражается в числах, то анализ множественной регрессии позволяет вам выяснить, какая именно комбинация параметров технологического процесса (X- переменные) позволяет добиться нужного результата (переменная Y).

б) *Хеджирование* (страхование) на рынке ценных бумаг подразумевает формирование портфеля ценных бумаг (чаще всего фьючерсов и опционов), который в максимальной степени учитывает риск тех или иных активов. Если, например, вы храните определенный запас товарно-материальных ценностей, следует позаботиться о хеджировании его риска. Банки используют контракты на казначейские фьючерсы и опционы для хеджирования риска потерь в результате изменения процентных ставок по их депозитным счетам и ссудам. Сельскохозяйственные отрасли используют хеджирование для снижения риска, связанного с флуктуациями цен на товары. Процесс выбора "хеджевого" портфеля можно осуществлять с помощью анализа множественной регрессии. Взяв за основу данные прошедшего периода, можно попытаться объяснять движение цен на ваши активы (Y) изменениями курса ценных бумаг (X_1, X_2 и т.д.) Соответствующие коэффициенты регрессии покажут, какой процент ценных бумаг того или иного вида следует включать в "хеджевой" портфель, чтобы как можно больше снизить риск. Таким образом, множественная регрессия будет использоваться для регулирования и управления риском, которому подвергаются ваши активы.

12.1. Интерпретация результатов множественной регрессии

Как будет выглядеть компьютерная распечатка результатов и как можно интерпретировать эти результаты? Прежде всего мы приведем краткий обзор входных данных и основных результатов. Более подробное их объяснение будет дано позже.

Пусть k означает количество поясняющих переменных (X-переменных); k может быть любым разумным числом. Ваши элементарные единицы нередко называются *наблюдениями*; это могут быть клиенты, фирмы, выпускаемые изде-

лия и т.п.[1] Входные данные для обычного множественного регрессионного анализа представлены в табл. 12.1.1

Сдвиг, или **постоянный член**, a, определяет прогнозируемое значение Y, когда *все* переменные X равны 0. **Коэффициент регрессии** для каждой X-переменной определяет влияние этой X-переменной на Y при условии, что все остальные X-переменные не меняются: коэффициент регрессии b_j для j-й X-переменной указывает, какое увеличение Y ожидается, когда все X-переменные остаются неизменными, за исключением переменной X_j, которая увеличивается на одну единицу. Взятые вместе эти коэффициенты регрессии составляют **уравнение прогнозирования**, или **уравнение регрессии**, вида (прогнозируемое значение Y) $= a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$, которое можно использовать в целях прогнозирования или управления. Эти коэффициенты (a, b_1, b_2, ... , b_k) обычно вычисляются методом наименьших квадратов, который минимизирует сумму квадратов ошибок прогнозирования. **Ошибки прогнозирования**, или **остатки**, определяются как Y – (прогнозируемое значение Y).

Как и в случае простой регрессии (с единственной X- переменной), **стандартная ошибка оценки**, S_e, указывает приблизительную величину ошибок прогнозирования. И как в случае простой регрессии, R^2 является **коэффициентом детерминации**, который указывает, какой процент вариации Y "объясняется" всеми X-переменными.[2]

Статистический вывод начинается с общей проверки, которую называют **F-тестом** (F-test). Цель F-теста заключается в том, чтобы выяснить, объясняют ли X-переменные значимую долю вариации Y. Если ваша регрессия *не* является значимой, говорить больше не о чем. Если же регрессия оказывается значимой, можно продолжить анализ статистических выводов, используя **t-тесты для отдельных коэффициентов регрессии**, которые показывают, насколько значимой является влияние той или иной X-переменной на Y *при условии, что все другие X-переменные остаются неизменными*. Построение доверительных интервалов и проверки гипотез для отдельного коэффициента регрессии будут, конечно же, основываться на его стандартной ошибке. Каждый коэффициент регрессии имеет свою стандартную ошибку; они обозначаются S_{b_1}, S_{b_2}, ... , S_{b_k}. В табл. 12.1.2 приведен перечень результатов множественного регрессионного анализа.

Пример. *Реклама в журналах*

Тарифы на размещение рекламных объявлений в журналах определяются каждым журналом самостоятельно. Чем объясняются различия в тарифах? Возможно, здесь каким-то образом учитывается ценность рекламного объявления для рекламодателя. Журналы, располагающие большей читательской аудиторией (при равных прочих условиях), наверное, вправе устанавливать большие тарифы. Кроме того, журналы, рассчитанные на более состоятельные круги читателей, также вправе устанавливать более высокие тарифы. Несмотря то что наверняка имеются и другие, не менее важные факторы, мы ограничимся лишь указанными двумя, добавив к ним еще один — предпочтения людей разного пола, и выясним, изменяют ли журналы свои тарифы в зависимости от соотношения мужчин и женщин в их читательской аудитории. Ответы на некоторые из этих вопросов можно получить с помощью множественного регрессионного анализа. Такой анализ поможет нам объяснить влияние на тарифы таких факторов, как величина читательской аудитории, структура читательской аудитории по полу и доходы читателей.

В табл. 12.1.3 представлена соответствующая многомерная совокупность данных, которую нам предстоит проанализировать. В качестве переменной Y (объясняемой) мы будем рассматривать стоимость одной страницы одноразовой полноцветной рекламы. Объясняющими переменными будут X_1, читательская аудитория (планируемая в тысячах человек), X_2, процент мужчин среди планируемой аудитории, и X_3, медиана дохода семьи. Размер выборки $n = 55$.

В табл. 12.1.4 представлена компьютерная распечатка результатов анализа множественной регрессии, полученная с помощью MINITAB®. Другие пакеты программного обеспечения для статистических расчетов позволяют получить в основном такую же базовую информацию. Например, с помощью Excel® также можно выполнить анализ множественной регрессии (найдите пункт Data Analysis (Анализ данных) в меню Tools (Сервис)[3] и выберите команду Regression (Регрессия)). На рис. 12.1.1,а показано диалоговое окно регрессии в Excel, а на рис. 12.1.1,б — результаты анализа множественной регрессии в Excel. Эти результаты мы будем интерпретировать в следующем разделе.

Таблица 12.1.1. Входные данные для множественной регрессии

	Y (зависимая, или объясняемая, переменная)	X_1 (первая независимая, или объясняющая, переменная)	X_2 (вторая независимая, или объясняющая, переменная)	...	X_k (последняя независимая, или объясняющая, переменная)
Наблюдение 1	10,9	2,0	4,7	...	12,5
Наблюдение 2	23,6	4,0	3,4	...	12,3
–	–	–	–	–	–
–	–	–	–	–	–
–	–	–	–	–	–
Наблюдение n	6,0	0,5	3,1	...	7,0

Таблица 12.1.2. Результаты множественного регрессионного анализа

Название	Результат	Описание
Сдвиг или постоянный член	a	Прогнозируемое значение для Y, когда все значения X-переменных равны 0
Коэффициенты регрессии	$b_1, b_2, \ldots, b_k$	Влияние каждой X-переменной на Y при условии, что все другие X-переменные остаются неизменными
Уравнение прогнозирования, или уравнение регрессии	Прогнозируемое значение $Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$	Прогнозируемое значение Y при заданных значениях X-переменных

[1] По "техническим" причинам у вас должно быть по крайней мере на одно наблюдение больше, чем имеется X-переменных, т.е. $n \geq k+1$. Практические соображения диктуют необходимость намного большего числа наблюдений.

[2] Однако в данном случае речь идет не просто о квадрате коэффициента корреляции Y с *одной* X-переменной, а о квадрате коэффициента корреляции r переменной Y (фактических значений) с прогнозами (которые вычисляются с помощью уравнения регрессии, найденного методом наименьших квадратов). Такой показатель учитывает *все* X-переменные.

[3] Если в меню Tools (Сервис) отсутствует пункт Data Analysis (Анализ данных), то сначала убедитесь, что вы выбрали ячейку электронной таблицы (а не график, например). Если вы все же не можете найти Data Analysis (Анализ данных), поищите пункт меню Add-Ins (Надстройки) и поставьте отметку возле Analysis ToolPak (Пакет анализа). Если это не поможет, то, видимо, необходимо переустановить Excel.

Название	Результат	Описание
Ошибки прогнозирования, или остатки	Y – прогнозируемое значение Y	Ошибка, возникающая для каждого наблюдения в результате использования уравнения прогнозирования вместо фактического значения Y для этого наблюдения
Стандартная ошибка оценки	S_e или S	Приблизительная величина ошибок прогнозирования (типичная разница между фактическим значением Y и его прогнозом исходя из уравнения регрессии)
Коэффициент детерминации	R^2	Процент изменчивости Y, объясняемый всей группой X-переменных
F-тест	Значимый или незначимый	Проверяет, может ли прогноз на основе X-переменных как группы быть лучше прогноза на основе простой случайности; по сути, проверяет, является ли R^2 большим, чем в случае отсутствия взаимосвязи между X-переменными и Y
t-тесты для отдельных коэффициентов регрессии	Значимый или незначимый, для каждой X-переменной	Проверяет, влияет ли на Y конкретная X-переменная при условии, что все другие X-переменные остаются неизменными; эту проверку выполняют только тогда, когда F-тест значим
Стандартные ошибки коэффициентов регрессии	S_{b_1}, S_{b_2}, ..., S_{b_k}	Указывает выборочную оценку стандартного отклонения каждого коэффициента регрессии; используется обычным способом для нахождения доверительных интервалов и проверки гипотез для отдельных коэффициентов регрессии
Число степеней свободы для стандартных ошибок коэффициентов регрессии	$n - k - 1$	Используется, чтобы найти в t-таблице соответствующее значение для построения доверительных интервалов и проверки гипотез для отдельных коэффициентов регрессии

Таблица 12.1.3. Тарифы на размещение рекламы и характеристики журналов

	Y, тариф (одна страница цветной рекламы), дол.	X_1, планируемая аудитория, тыс. человек	X_2, процент мужчин	X_3, медиана дохода семьи, дол.
Audubon	25 315	1 645	51,1	38 787
Better Homes & Gardens	198 000	34 797	22,1	41 933
Business Week	103 300	4 760	68,1	63 667
Cosmopolitan	94 100	15 452	17,3	44 237
Elle	55 540	3 735	12,5	47 211
Entrepreneur	40 355	2 476	60,4	47 579
Esquire	51 559	3 037	71,3	44 715
Family Circle	147 500	24 539	13,0	38 759
First For Women	28 059	3 856	3,6	43 850
Forbes	59 340	4 191	68,8	66 606
Fortune	60 800	3 891	68,8	58 402
Glamour	85 080	10 891	7,8	46 331
Golf Digest	98 760	6 250	78,9	61 323

	Y, тариф (одна страница цветной рекламы), дол.	X_1, планируемая аудитория, тыс. человек	X_2, Процент мужчин	X_3, Медиана дохода семьи, дол.
Good Housekeeping	166 080	25 306	12,6	38 335
Gourmet	49 640	4 484	29,6	57 060
Harper's Bazaar	52 805	2 621	11,5	44 992
Inc.	70 825	2 166	66,9	72 493
Kiplinger's Personal Finance	46 580	3 332	65,1	63 876
Ladies' Home Journal	127 000	17 040	6,8	38 442
Life	63 750	14 220	46,9	41 770
Mademoiselle	55 910	4 804	8,0	46 694
Martha Stewart's Living	93 328	4 849	16,6	61 890
McCalls	113 120	16 301	7,6	33 823
Money	98 250	9 805	60,6	60 549
Motor Trend	79 800	5 281	88,5	48 739
National Geographic	159 345	32 158	53,0	44 326
Natural History	20 180	1 775	45,0	41 499
Newsweek	148 800	20 720	53,5	53 025
Parents Magazine	72 820	12 064	18,2	39 369
PC Computing	40 675	4 606	67,0	57 916
People	125 000	33 668	34,0	46 171
Popular Mechanics	78 685	9 036	86,9	40 802
Reader's Digest	193 000	51 925	42,4	38 060
Redbook	95 785	13 212	8,9	41 156
Rolling Stone	78 920	8 638	59,8	43 212
Runner's World	36 850	2 078	62,9	60 222
Scientific American	37 500	2 704	70,0	62 372
Seventeen	71 115	5 738	17,0	37 034
Ski	32 480	2 249	64,5	58 629
Smart Money	42 900	2 224	63,4	57 170
Smithsonian	73 075	8 253	47,9	50 872
Soap Opera Digest	35 070	7 227	10,3	31 835
Sports Illustrated	162 000	21 602	78,8	45 897
Sunset	56 000	5 276	38,7	52 524
Teen	53 250	3 057	15,4	42 640
The New Yorker	62 435	3 223	48,9	49 672
Time	162 000	22 798	52,4	49 166

Окончание табл. 12.1.3

	Y, тариф (одна страница цветной рекламы), дол.	X_1, планируемая аудитория, тыс. человек	X_2, Процент мужчин	X_3, Медиана дохода семьи, дол.
True Story	17 100	3 582	12,2	15 734
TV Guide	146 400	40 917	42,8	37 396
U.S. News & World Report	98 644	9 825	57,5	52 018
Vanity Fair	67 890	4 307	27,7	52 189
Vogue	63 900	8 434	12,9	44 242
Woman's Day	137 000	22 747	6,7	38 463
Working Woman	87 500	3 312	6,3	44 674
YM	73 270	3 109	14,4	43 696
Среднее значение	83 534	10 913	39,7	47 710
Среднеквадратическое отклонение	45 446	11 212	25,9	10 225

Размер выборки: $n = 55$.

Данные взяты из *Mediamark Research Magazine Qualitative Audiences Report*, Spring 1996; и *SDRS Consumer Magazine Advertising Source*, July 1997, Volume 79 Number 7.

Коэффициенты регрессии и уравнение регрессии

Сдвиг, или постоянный член, a, и коэффициенты регрессии, b_1, b_2 и b_3, вычисляются компьютером с использованием метода наименьших квадратов. Среди всех возможных вариантов уравнения регрессии с различными значениями этих коэффициентов именно уравнение, найденное таким методом, обеспечивает минимальную сумму квадратов ошибок прогнозирования для рассматриваемой нами выборки журналов. Уравнение регрессии (или уравнение прогнозирования) имеет следующий вид:

(прогнозируемый тариф на размещение рекламы) $= a + b_1X_1 + b_2X_2 + b_3X_3$
$=$ \$4 043 + 3,79(читательская аудитория) − 124(процент мужчин) + 0,903(медиана дохода).

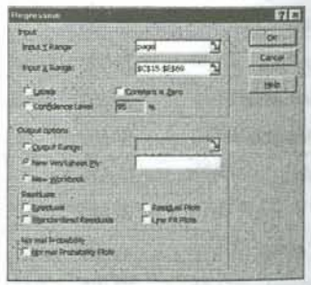
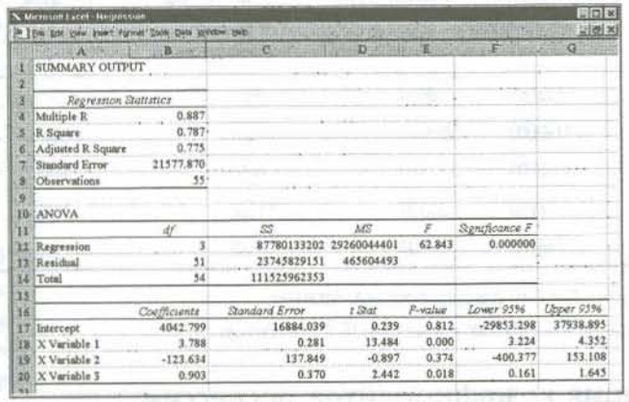


Рис. 12.1.1.а) Диалоговое окно регрессии в Excel®. Можно присвоить имя диапазона для Y (в данном случае — "page"), но X-переменные должны находиться в смежных столбцах: можно протащить мышь по столбцам (только данные, без названий над ними) или ввести адрес соответствующей ячейки

Сдвиг, a = \$4 043, интерпретируется следующим образом: типичный тариф на размещение одностраничного цветного рекламного объявления в журнале, у которого нет платных подписчиков, нет мужчин среди читателей и читатели не имеют дохода, составляет \$4 043. Однако в рассматриваемой нами совокупности данных нет подобных журналов, поэтому сдвиг, a, следует рассматривать лишь как вспомогательную величину, необходимую для получения оптимальных прогнозов, но не интерпретировать это значение так буквально.



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.887					
5	R Square	0.787					
6	Adjusted R Square	0.775					
7	Standard Error	21577.870					
8	Observations	55					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	87780133202	29260044401	62.843	0.000000	
13	Residual	51	23745829151	465604493			
14	Total	54	111525962353				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	4042.799	16884.039	0.239	0.812	-29853.298	37938.895
18	X Variable 1	3.788	0.281	13.484	0.000	3.224	4.352
19	X Variable 2	-123.634	137.849	-0.897	0.374	-400.377	153.108
20	X Variable 3	0.903	0.370	2.442	0.018	0.161	1.645

Рис. 12.1.1.б) Полученные в Excel результаты регрессионного анализа данных о рекламных объявлениях в журналах

Таблица 12.1.4. Результат множественного регрессионного анализа тарифов на размещение рекламы в журналах (вычисления сделаны компьютерным пакетом программ MINITAB)

Уравнение регрессии имеет вид

тариф на размещение рекламы = 4 043 + 3,79 (аудитория) − 124 (процент мужчин) + 0,903 (доход)

Независимая переменная	Коэффициент	Стандартное отклонение	t	p
Константа	4043	16884	0,24	0,812
Аудитория	3,7880	0,2809	13,48	0,000
Процент мужчин	−123,6	137,8	−0,90	0,374
Доход	0,9026	0,3696	2,44	0,018

S = 21578 R-квадрат = 78,7% R-квадрат(кор.) = 77,5%

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	3	87780133202	29260044401	62,84	0,000
Ошибка остатка	51	23745829151	465604493		
Итого	54	1,11526E+11			

Источник	DF	Seq SS
Аудитория	1	84858244860
Процент мужчин	1	144950723
Доход	1	2776937619

Окончание табл. 12.1.4

Необычные наблюдения

Наблюдения	Аудитория	Тариф на рекламу	Соответствие	Стандартное отклонение соответствия	Остаток	Стандартное отклонение остатка
31	33668	125000	169049	6939	−44049	−2,16R
33	51925	193000	229848	11268	−36848	−2,00RX
43	21602	162000	117556	6850	44444	2,17R
48	3582	17100	30305	11490	−13205	−0,72X
49	40917	146400	187500	8582	−41100	−2,08R

R обозначает наблюдение с большим стандартизованным остатком;

X – наблюдение, X-значение которого обеспечивает ему большое влияние.

Интерпретация коэффициентов регрессии

Коэффициенты регрессии интерпретируются как влияние каждой из переменных на размер тарифа, если все другие независимые ("объясняющие") переменные остаются неизменными. Часто это значение включает "поправку на" другие независимые переменные, или "контролирование" этих других независимых переменных. Поэтому коэффициент регрессии для конкретной X-переменной может изменяться (иногда значительно) в результате включения в анализ или исключения других X-переменных. В частности, каждый коэффициент регрессии определяет среднее увеличение тарифа на размещение рекламы, приходящееся на единичное увеличение соответствующей ему X-переменной (в данном случае термин "единичное" означает одну единицу измерения конкретной X-переменной).

Коэффициент регрессии для размера читательской аудитории, $b_1 = 3{,}79$, указывает, что — *при всех прочих равных условиях* — журнал с дополнительной тысячью читателей (поскольку у нас X_1 измеряется в тысячах человек) берет (в среднем) на \$3,79 больше за размещение одностраничного цветного рекламного объявления. Можно также считать, что коэффициент регрессии для размера читательской аудитории означает, что каждый дополнительный читатель увеличивает для этого журнала тариф на размещение рекламных объявлений на \$0,00379, т.е. увеличение составляет чуть меньше половины цента на одного человека. Поэтому, если у какого-то другого журнала такой же процент читателей-мужчин и такой же показатель медианы дохода семьи читателей, но читательская аудитория на 3548 человек больше, то можно ожидать, что тариф на размещение рекламных объявлений в этом журнале будет (в среднем) на $3{,}79 \times 3{,}548 = \$13{,}45$ больше благодаря такому отличию размера читательской аудитории.

Коэффициент регрессии для процента мужчин, $b_2 = -124$, указывает, что (при всех прочих равных условиях) тариф на размещение цветных рекламных объявлений в журнале с дополнительным 1% читателей-мужчин окажется (в среднем) на \$124 меньше. Это означает, что читательницы представляют для журнала большую ценность, чем читатели-мужчины. Статистический вывод должен подтвердить или опровергнуть эту гипотезу путем сравнения величины влияния процента мужчин (т.е. −\$124) с тем, на что можно было бы рассчитывать, если бы при данных обстоятельствах все определялось лишь чистой случайностью.

Коэффициент регрессии для медианы дохода, $b_3 = 0{,}903$, указывает, что (при всех прочих равных условиях) в журнале с дополнительным долларом медианы дохода его читателей тариф на размещение одностраничного цветного рекламного объявления будет (в среднем) на \$0,903 больше. Положительный знак этого коэффициента совершенно оправдан, поскольку люди с более высоким уровнем доходов могут позволить себе тратить больше на покупку рекламируемой продукции. Если у какого-то другого журнала такой же процент читателей-мужчин и такая же величина читательской аудитории, но медиана дохода семей читателей на \$4 000 выше, то можно ожидать, что тариф этого журнала на размещение рекламных объявлений будет на $0{,}903 \times 4000 = \$3612$ выше (в среднем) благодаря более высокому уровню доходов его читателей.

Помните, что коэффициенты регрессии отражают влияние на Y одной X-переменной при условии, что все другие X-переменные остаются неизменными. Это следует понимать буквально. Например, коэффициент регрессии b_3 отражает влияние медианы дохода читателей на рекламные тарифы; он вычисляется при неизменных величинах читательской аудитории и процента читателей-мужчин. В таком случае более высокие уровни доходов читателей, как правило, ведут к установлению более высоких тарифов на размещение рекламных объявлений (поскольку b_3 является положительным числом) — при фиксированных размере читательской аудитории и проценте читателей-мужчин.

Какой была бы эта взаимосвязь, если бы остальные переменные (размер читательской аудитории и процент читателей-мужчин) *не* фиксировались на постоянном уровне? На этот вопрос можно ответить, проанализировав обычный коэффициент корреляции (или коэффициент регрессии, прогнозирующий Y на основании только одной этой X-переменной), вычисленный только для двух переменных: тарифа и медианы дохода. В нашем случае более высокое значение медианы дохода фактически ассоциируется с *более низким* тарифом (корреляция тарифа и медианы дохода является отрицательной: −0,167)! Чем это объяснить? Вполне приемлемое объяснение заключается в том, что журналы, ориентирующиеся на читателей с более высоким средним уровнем доходов, не в состоянии обеспечить себе массовую аудиторию из-за того, что богатых людей среди населения страны в целом не так уж много. Если же эта читательская аудитория богатых людей окажется очень небольшой, это может вообще исказить эффект влияния высокого уровня доходов в расчете на одного читателя.

Прогнозы и ошибки прогнозирования

Уравнение прогнозирования, или уравнение регрессии, определяется в следующем виде:

$$(\text{прогнозируемое значение } Y) = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k.$$

В нашем примере с рекламными объявлениями в журналах, чтобы найти прогнозируемую величину тарифа на размещение рекламных объявлений исходя из величины читательской аудитории, процента читателей-мужчин и медианы

дохода читателей для конкретного журнала, подобного тем, которые составляют рассматриваемую нами выборку данных, подставим в уравнение прогнозирования соответствующие этому журналу значения X-переменных:

(прогнозируемый тариф на размещение рекламы) =
$= a + b_1X_1 + b_2X_2 + b_3X_3 = \$4\,043 + 3{,}79X_1 - 124X_2 + 0{,}903\,X_3 =$
= \$4 043 + 3,79(читательская аудитория) − 124(процент мужчин) +
+ 0,903(медиана дохода).

Допустим, например, что вы собираетесь основать новый журнал, *Популярная статистика*, который рассчитан на читательскую аудиторию порядка 900 000 человек, 55% которых будут составлять женщины, а медиана дохода его читателей равна \$50 000. Данные в уравнение прогнозирования необходимо подставить в той же форме, что и в исходной совокупности данных (т.е. той, исходя из которой и строилось уравнение регрессии): $X_1 = 900$ (читательская аудитория в тысячах человек), $X_2 = 45$ (процент мужчин) и $X_3 = \$50\,000$ (медиана дохода). Прогнозируемое значение для этой ситуации определяется следующим образом:

прогнозируемый тариф на размещение рекламы в журнале
Популярная статистика = 4 043 + 3,79(читательская аудитория) −
− 124(процент мужчин) + 0,903(медиана дохода) = 4 043 + 3,79 ×
× 900 − 124 × 45 + 0,903 × 50 000 = \$47 024.

Разумеется, рассчитывать на то, что тариф на размещение рекламы в журнале составит ровно \$47 024, не приходится. Во-первых, даже между журналами, данными о которых мы располагаем, наблюдаются случайные колебания, поэтому прогнозы не являются идеальными даже для них. Во-вторых, прогнозы могут быть полезны лишь в той мере, в какой прогнозируемый журнал подобен журналам, принадлежащим к исходной совокупности данных. Если речь идет о новом журнале, то тариф на размещение рекламы в этом журнале может определяться не так, как для журналов с уже устоявшейся репутацией, которые мы использовали для построения уравнения регрессии.

С помощью этого уравнения можно также прогнозировать тарифы для журналов, принадлежащих к исходной совокупности данных. У первого журнала, *Audubon*, $X_1 = 1\,645$ (читательская аудитория равна примерно 1,6 миллиона человек), $X_2 = 51{,}1$ (т.е. 51,1% читателей этого журнала — мужчины) и $X_3 = 38\,787$ (медиана годового дохода читателей этого журнала составляет \$38 787). Прогнозируемое значение для этого журнала можно найти по следующей формуле:

прогнозируемый тариф на размещение рекламы в журнале *Audubon* =
= 4 043 + 3,79(читательская аудитория) − 124(процент мужчин) +
+ 0,903(медиана дохода) = 4 043 + 3,79 × 1 645 − 124 × 51,1 +
+ 0,903 × 38 787 = \$38 966.

Остаток, или ошибка прогнозирования, определяется по формуле: Y − (прогнозируемое значение Y). Для журнала, принадлежащего к исходной совокупности данных, этот показатель равняется фактическому тарифу минус прогнозируемый тариф. Для журнала *Audubon* фактический тариф составляет

\$25 315, а прогнозируемый тариф — \$38 966. Таким образом, ошибка прогнозирования равна 25 315 − 38 966 = −\$13 651. Отрицательный остаток указывает на то, что фактический тариф меньше прогнозируемого (в случае журнала *Audubon* примерно на \$14 000). Для многих из нас \$14 000 — огромные деньги; неплохо бы взглянуть на другие ошибки прогнозирования, чтобы понять, в какой мере прогнозирование отражает реальную ситуацию. Почему рекламные тарифы в журнале *Audubon* оказались намного меньше их ожидаемой величины? Скорее всего, потому, что для прогнозирования использовалось лишь $k = 3$ из множества возможных факторов, влияющих на величину рекламных тарифов (к тому же многие из этих факторов не очень понятны и их довольно сложно измерить).

В табл. 12.1.5 показаны фактические тарифы и прогнозируемые тарифы (которые также называют *ожидаемыми*, или *подогнанными*, значениями), а также ошибки прогнозирования для каждого из журналов в исходной совокупности данных.

Таблица 12.1.5. Прогнозируемые значения и остатки для тарифов на размещение рекламы в журналах

	Тариф на размещение рекламы (фактический), дол.	Тариф на размещение рекламы (прогнозируемый), дол.	Ошибки прогнозирования (остатки)
Audubon	25 315	38 966	−13 651
Better Homes & Gardens	198 000	170 972	27 028
Business Week	103 300	71 120	32 180
Cosmopolitan	94 100	100 365	−6 265
Elle	55 540	59 258	−3 718
Entrepreneur	40 355	48 899	−8 544
Esquire	51 559	47 092	4 467
Family Circle	147 500	130 374	17 126
First For Women	28 059	57 783	−29 724
Forbes	59 340	71 531	−12 191
Fortune	60 800	62 990	−2 190
Glamour	85 080	86 152	−1 072
Golf Digest	98 760	73 314	25 446
Good Housekeeping	166 080	132 946	33 134
Gourmet	49 640	68 871	−19 231
Harper's Bazaar	52 805	53 159	−354
Inc.	70 825	69 409	1 416
Kiplinger's Personal Finance	46 580	66 271	−19 691
Ladies' Home Journal	127 000	102 448	24 552
Life	63 750	89 812	−26 062
Mademoiselle	55 910	63 398	−7 488

Окончание табл. 12.1.5

	Тариф на размещение рекламы (фактический), дол.	Тариф на размещение рекламы (прогнозируемый), дол.	Ошибки прогнозирования (остатки)
Martha Stewart's Living	93 328	76 221	17 107
McCalls	113 120	95 381	17 739
Money	98 250	88 344	9 906
Motor Trend	79 800	57 098	22 702
National Geographic	159 345	159 315	30
Natural History	20 180	42 660	–22 480
Newsweek	148 800	123 777	25 023
Parents Magazine	72 820	83 026	–10 206
PC Computing	40 675	65 482	–24 807
People	125 000	169 049	–44 049
Popular Mechanics	78 685	64 356	14 329
Reader's Digest	193 000	229 848	–36 848
Redbook	95 785	90 138	5 647
Rolling Stone	78 920	68 374	10 546
Runner's World	36 850	58 494	–21 644
Scientific American	37 500	61 928	–24 428
Seventeen	71 115	57 104	14 011
Ski	32 480	57 506	–25 026
Smart Money	42 900	56 231	–13 331
Smithsonian	73 075	75 301	–2 226
Soap Opera Digest	35 070	58 880	–23 810
Sports Illustrated	162 000	117 556	44 444
Sunset	56 000	66 652	–10 652
Teen	53 250	52 206	1 044
The New Yorker	62 435	55 040	7 395
Time	162 000	128 301	33 699
True Story	17 100	30 305	–13 205
TV Guide	146 400	187 500	–41 100
U.S. News & World Report	98 644	81 103	17 541
Vanity Fair	67 890	64 039	3 851
Vogue	63 900	74 329	–10 429
Woman's Day	137 000	124 098	12 902
Working Woman	87 500	56 133	31 367
YM	73 270	53 480	19 790

Насколько хороши наши прогнозы

Этот раздел следует рассматривать в основном как обзор, поскольку стандартное отклонение оценки, S_e, и коэффициент детерминации, R^2, имеют для множественной регрессии, вообще говоря, ту же интерпретацию, что и для простой регрессии, речь о которой шла в предыдущей главе. Единственное отличие заключается в том, что ваши прогнозы теперь базируются на нескольких X-переменных. Но все остается очень похоже, поскольку вы по-прежнему прогнозируете только одну переменную Y.

Типичная ошибка прогнозирования: стандартная ошибка оценки

Как и в случае простой регрессии, когда мы имеем дело лишь с одной X-переменной, стандартная ошибка оценки указывает приблизительную величину ошибок прогнозирования. Возвращаясь к нашему примеру с тарифами на размещение рекламы в журналах, S_e = \$21 578. Это говорит о том, что фактические тарифы на размещение рекламы в этих журналах, как правило, отклоняются от прогнозируемых тарифов не более чем на \$21 578 (речь идет о стандартном отклонении). Иными словами, если распределение ошибок является нормальным, то можно ожидать, что примерно 2/3 фактических тарифов будут находиться в пределах S_e от прогнозируемых тарифов; примерно 95% — в пределах $2S_e$ и т.д.

Эта стандартная ошибка оценки, S_e = \$21 578, указывает остаток вариации тарифов после того, как вы использовали X-переменные (величина читательской аудитории, процент мужчин и медиана дохода) в уравнении регрессии для прогнозирования тарифов каждого журнала. Сравните этот показатель с обычным стандартным отклонением одной переменной для тарифов, S_Y = \$45 446, вычисленным без использования других переменных. Это стандартное отклонение, S_Y, указывает остаток вариации тарифов после того, как вы использовали для прогнозирования тарифов каждого журнала только значение $\bar{Y}$. Заметьте, что S_e = \$21 578 меньше, чем S_Y = \$45 446; ошибки, как правило, оказываются меньше, если для прогнозирования тарифов использовать уравнение регрессии, а не просто $\bar{Y}$. Как видите, X-переменные полезны для объяснения размеров тарифов.

Это можно представить себе следующим образом. Если вам ничего неизвестно об X-переменных, вы будете использовать в качестве оптимальной приблизительной оценки среднее значение тарифа ($\bar{Y}$ = \$83 534) и будете ошибаться приблизительно на S_Y = \$45 446. Но если вам известны такие характеристики, как величина читательской аудитории, процент мужчин и средний доход, то для прогнозирования тарифов можно воспользоваться уравнением регрессии; в этом случае вы ошибетесь примерно на S_e = \$21 578. Такое сокращение ошибки прогнозирования (с \$45 446 до \$21 578) и является одним из преимуществ использования регрессионного анализа.

Объясненный процент вариации: R^2

Коэффициент детерминации (часто также используют термин "квадрат множественной корреляции". — *Прим. ред.*), R^2, указывает, какой процент вариации Y объясняется влиянием всех X-переменных.

Если вернуться к нашему примеру с тарифами на размещение рекламы в журналах, то коэффициент детерминации, $R^2 = 0{,}787$, или 78,7%, указывает на то, что независимые переменные (X-переменные величины читательской аудитории, процент мужчин и средний доход) объясняют 78,7% вариации тарифов.[4] При этом 21,3% остаются необъясненными и связываются с влиянием других факторов. 78,7% — довольно большое значение R^2; во многих исследованиях приходится работать со значительно меньшими величинами, которые, тем не менее, обеспечивают достаточно качественные прогнозы. Желательно, чтобы значение R^2 было как можно большим (большие значения R^2 свидетельствуют о том, что исследуемая взаимосвязь является достаточно сильной). В идеальном случае $R^2 = 100\%$; это возможно лишь в том случае, когда все ошибки прогнозирования равны 0 (что, как правило, свидетельствует о наличии ошибок в другом месте!).

Статистический вывод в случае множественной регрессии

Полученные нами к настоящему времени результаты регрессии представляют собой достаточно полное описание исследуемых ($n = 55$) журналов, однако статистический вывод помог бы нам обобщить этот случай на идеализированную популяцию подобных им журналов. Вместо того чтобы просто констатировать тот факт, что увеличение на один процент числа читателей-мужчин приводит к уменьшению тарифа на размещение рекламы в среднем на \$124, можно сделать статистический вывод относительно большой генеральной совокупности журналов такого типа, из которой вполне могли бы быть извлечены имеющиеся данные, и попытаться выяснить, существует ли *в действительности* какая-либо взаимосвязь между полом читателей журнала и тарифами на рекламу, или коэффициент регрессии, равный −\$124, можно объяснить просто случайностью. Может ли быть так, что обнаруженное нами влияние процента читателей-мужчин на стоимость рекламы — это просто случайное число, а не свидетельство наличия систематической взаимосвязи? Ответ на этот вопрос можно получить с помощью статистического вывода.

В табл. 12.1.6 содержится часть результатов работы компьютерной программы, приведенных в табл. 12.1.4. Здесь статистические выводы можно делать на основе p-значений как для общего F-теста, так и для тестов относительно каждой из независимых X-переменных. Мы подробно обсудим все это в последующих разделах — после определения генеральной совокупности, относительно которой мы собираемся сделать статистический вывод.

Предположения

Чтобы не усложнять пример, предположим, что мы располагаем случайной выборкой из намного большей генеральной совокупности. Допустим также, что эта генеральная совокупность характеризуется линейной взаимосвязью со случайностью, представленной **моделью множественной линейной регрессии**, в соответствии с которой наблюдаемое значение Y определяется взаимосвязью в генеральной совокупности плюс нормально распределенная случайная ошибка. Предполагается также, что эти случайные ошибки для разных наблюдений (элементарных единиц наших данных) не зависят друг от друга.

[4] С технической точки зрения это та часть *дисперсии* (квадрат стандартного отклонения) Y, которая объясняется X-переменными.

Модель множественной линейной регрессии для генеральной совокупности

$$Y = (\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k) + \varepsilon =$$

$$= (\text{взаимосвязь в генеральной совокупности}) + \text{случайность},$$

где ε характеризуется нормальным распределением со средним значением 0 и постоянным стандартным отклонением σ, причем эта случайность является независимой для разных наблюдений (элементарных единиц данных).

Взаимосвязь в генеральной совокупности определяется $k + 1$ параметрами: α представляет сдвиг (или постоянный член) для генеральной совокупности, а β_1, β_2, ... ,β_k являются коэффициентами регрессии для генеральной совокупности, которые показывают среднее влияние каждой из X-переменных на Y (в данной генеральной совокупности), при условии, что все остальные X-переменные остаются неизменными. Сводка параметров генеральной совокупности и характеристик выборки приведена в табл. 12.1.7. Если бы вы имели данные обо всей генеральной совокупности, то полученные вами с помощью метода наименьших квадратов коэффициенты регрессии ничем не отличались бы от соответствующих коэффициентов, описывающих связь в генеральной совокупности. Как правило, однако, полученный методом наименьших квадратов сдвиг a является лишь *статистической оценкой* α, а полученные методом наименьших квадратов коэффициенты регрессии b_1, b_2, ... ,b_k представляют лишь *статистические оценки* β_1, β_2, ... ,β_k соответственно. Существуют, конечно же, ошибки, обусловленные процессом оценивания, поскольку выборка намного меньше всей генеральной совокупности.

Таблица 12.1.6. Статистический вывод для тарифов на размещение рекламы в журналах

Независимая переменная	Коэффициент	Стандартное отклонение	t	p
Константа	4043	16884	0,24	0,812
Аудитория	3,7880	0,2809	13,48	0,000
Процент мужчин	−123,6	137,8	−0,90	0,374
Доход	0,9026	0,3696	2,44	0,018

S = 21578; R-квадрат = 78,7% R-квадрат(кор.) = 77,5%.

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	3	87780133202	29260044401	62,84	0,000
Ошибка остатка	51	23745829151	465604493		
Итого	54	1,11526E+11			

Таблица 12.1.7. Параметры генеральной совокупности и характеристики выборки для множественной регрессии

	Генеральная совокупность (параметры: фиксированные и неизвестные)	Выборка (статистические оценки: случайные и известные)
Сдвиг, или постоянный член	α	a
Коэффициенты регрессии	β_1	b_1
	β_2	b_2
	–	–
	–	–
	–	–
	β_k	b_k
Неопределенность Y	σ	S_e

Как на диаграмме рассеяния представить множественную линейную регрессионную взаимосвязь? Каждый раз, когда добавляется новая независимая переменная X, *добавляется еще одно измерение. Например, при наличии лишь одной* X-переменной (см. главу 11) мы имели линию прогнозирования в плоском, двумерном пространстве. При наличии двух X-переменных можно говорить о плоскости прогнозирования в трехмерном пространстве с измерениями X_1, X_2 и Y, как показано на рис. 12.1.2. Одно из предположений множественного регрессионного анализа заключается в том, что взаимосвязь в генеральной совокупности является, по существу, плоской, а не изогнутой.

Значима ли модель? F-тест или тест R^2

Статистический вывод начинается с F-теста, целью которого является выяснение, объясняют ли X-переменные значимую часть вариации Y. F-тест используется как "входные ворота" в статистический вывод: если этот тест значим, следовательно, связь существует и можно приступать к ее исследованию и объяснению. Если этот тест незначим, то мы имеем дело с набором не связанных между собой случайных чисел — объяснять, в сущности, нечего. Помните, что, когда вы принимаете нулевую гипотезу, это считается *слабым* заключением. Вы не доказали, что взаимосвязи нет: вам просто не хватает убедительных доводов в пользу наличия такой взаимосвязи. Взаимосвязь вполне может существовать, но из-за случайности или малого размера выборки вы не в состоянии обнаружить ее с помощью тех данных, которые имеются в вашем распоряжении.

Нулевая гипотеза для F-теста утверждает, что в генеральной совокупности между X-переменными и Y прогнозирующая взаимосвязь *отсутствует*. Иначе говоря, Y является чисто случайной величиной и значения X-переменных не оказывают на Y никакого влияния. Если посмотреть на модель множественной линейной регрессии, то это утверждение означает, что $Y = \alpha + \varepsilon$, что может иметь место в том случае, если *все* коэффициенты регрессии в генеральной совокупности равны 0.

Альтернативная гипотеза F-теста утверждает, что в генеральной совокупности между X-переменными и Y существует определенная прогнозирующая взаимосвязь. Таким образом, переменная Y уже не является чисто случайной величиной и должна зависеть по крайней мере от одной из X-переменных. Иными словами, альтернативная гипотеза утверждает, что *по крайней мере один* из коэффициентов регрессии не равен 0. Обратите внимание: вовсе не обязательно, чтобы каждая из X-переменных влияла на Y — достаточно, чтобы влияла хотя бы одна из них.

Гипотезы для F-теста

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

H_1: по крайней мере один из $\beta_1, \beta_2, \ldots, \beta_k \neq 0$.

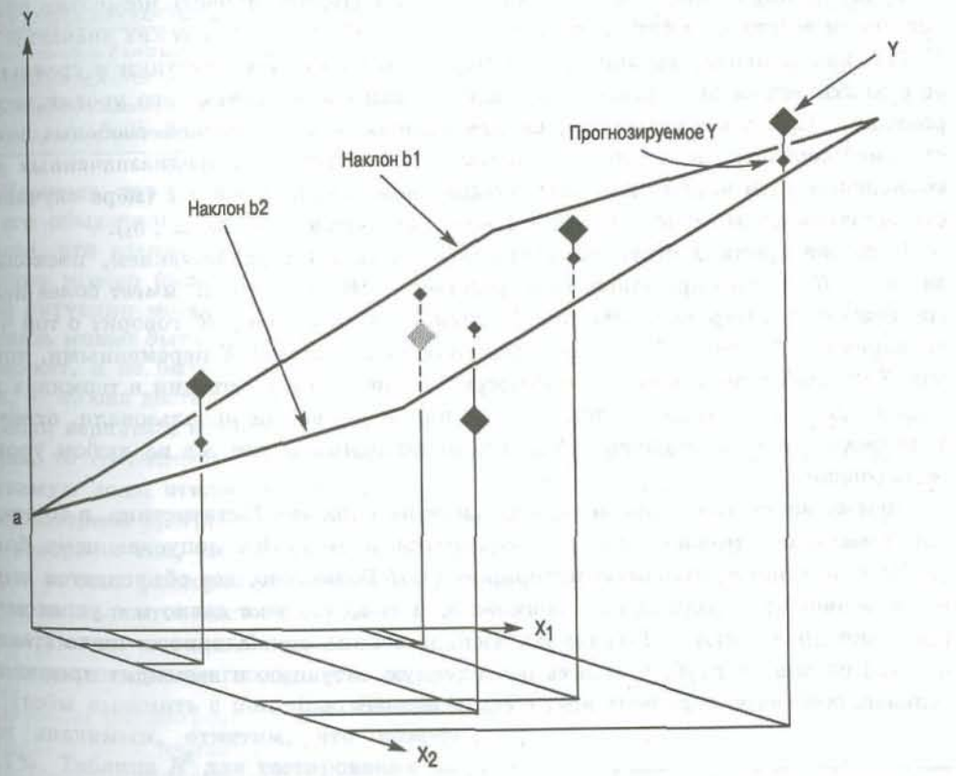


Рис. 12.1.2. Когда две независимые X-переменные используются для прогнозирования Y, уравнение прогнозирования можно представить в виде плоскости, ближайшей к точкам данных в трехмерном пространстве. Сдвиг a определяется точкой, в которой эта плоскость прогнозирования пересекает ось Y. Коэффициенты регрессии b_1 и b_2 определяют наклон плоскости прогнозирования в двух соответствующих направлениях

Выполнить F-тест проще всего, отыскав в результатах работы компьютерной программы подходящее p-значение и интерпретировав результирующий уровень значимости, как мы делали это в главе 10. Если p-значение больше, чем 0,05, то полученный результат не является значимым. Если же это p-значение меньше, чем 0,05, то полученный результат является значимым. Если $p < 0{,}01$, тогда полученный результат является высоко значимым, и т.д.

Еще один способ выполнения F-теста заключается в сравнении значения R^2 (процент вариации Y, который объясняется X-переменными) со значениями из таблицы критических значений R^2 для подходящего уровня тестирования (например, 5%). Если значение R^2 оказывается достаточно большим, тогда регрессия считается значимой, т.е. удалось объяснить больше, чем просто случайную величину вариации Y. Эта таблица индексирована по n (количество наблюдений) и k (количество X-переменных).

Традиционный способ выполнения F-теста интерпретировать несколько сложнее, но он всегда дает тот же результат, что и таблица критических значений R^2. F-тест, как правило, выполняется путем вычисления F-статистики и сравнения ее с критическим значением из F-таблицы для соответствующего уровня тестирования.[5] При этом используются два разных числа степеней свободы: число степеней свободы числителя k (количество X-переменных, предназначенных для объяснения Y) и число степеней свободы знаменателя $n - k - 1$ (мера случайности остатков *после* оценивания $k + 1$ коэффициентов a, b_1, b_2, ... , b_k).

В то же время F-статистика является излишним усложнением, поскольку значение R^2 можно проверить непосредственно. Более того, R^2 имеет более непосредственную интерпретацию, чем F-статистика, поскольку R^2 говорит о той части вариации Y, которая учитывается (или объясняется) X-переменными, тогда как F не имеет столь простой и непосредственной интерпретации в терминах исходных данных. Какой бы подход — F или R^2 — вы ни использовали, ответ (о значимости или не значимости) всегда будет одним и тем же на любом уровне тестирования.

Почему же по традиции используется более сложная F-статистика, в то время как вместо нее можно было бы обратиться к тесту R^2, допускающему более удобную и непосредственную интерпретацию? Возможно, все объясняется именно сложившейся традицией, а возможно, и тем, что уже давно и с успехом на практике применяются F-таблицы. Использование осмысленного числа (такого как R^2) позволяет глубже понять исследуемую ситуацию и выглядит предпочтительнее, особенно когда речь идет о сфере бизнеса.

[5] Для особо интересующихся заметим, что F-статистика получила свое название в честь сэра Рональда А. Фишера и определяется как "объясненное среднеквадратическое", деленное на "необъясненное среднеквадратическое". Большие значения F предполагают, что регрессионная модель является значимой, поскольку удалось объяснить довольно значительную долю вариации Y в сравнении с долей необъясненной случайности. Большие значения R^2 также предполагают значимость. Связь между F и R^2 состоит в том, что $F = (n - k - 1)[1/(1 - R^2) - 1]/k$, а $R^2 = 1 - 1/[1 + kF/(n - k - 1)]$, а значит, большим значениям F соответствуют большие значения R^2 (и наоборот). Вот почему тесты на большие значения F полностью соответствуют тестам на большие значения R^2.

Результат F-теста (решение принимается на основе p-значения)

Если p-значение *больше*, чем 0,05, значит, соответствующая модель *не является значимой* (вы принимаете нулевую гипотезу о том, что X-переменные *не* помогают прогнозировать Y).

Если p-значение оказывается *меньше*, чем 0,05, значит, соответствующая модель *является значимой* (вы отвергаете нулевую гипотезу и принимаете альтернативную гипотезу о том, что X-переменные помогают прогнозировать Y).

Результат F-теста (решение принимается на основе R^2)

Если значение R^2 *меньше*, чем критическое значение в таблице R^2, значит, соответствующая модель *не является значимой*. Если значение R^2 *больше*, чем критическое значение в таблице R^2, значит, соответствующая модель *является значимой*. Этот ответ в любом случае будет таким же, как результат, полученный с помощью p-значения.

Результат F-теста (решение принимается непосредственно на основе F)

Если значение F оказывается *меньше*, чем критическое значение в F-таблице, значит, соответствующая модель *не является значимой*. Если значение F оказывается *больше*, чем критическое значение в F-таблице, — соответствующая модель *является значимой*. Этот ответ в любом случае будет таким же, как результат, полученный с помощью p-значения или R^2.

Помните, что статистический смысл термина *значимый* несколько отличается от его обыденного смысла. Когда вы находите значимую модель регрессии, то знаете, что взаимосвязь между X-переменными и Y оказывается сильнее, чем обычно можно было бы ожидать от чистой случайности. Другими словами, в этой ситуации можно говорить о наличии определенной взаимосвязи. Эта взаимосвязь может быть сильной или полезной в том или ином практическом смысле (а может, и не быть таковой) — эти вопросы требуют специального рассмотрения, — но она достаточно сильна, чтобы не выглядеть как чистая случайность.

Если вернуться к нашему примеру с тарифами на размещение рекламы в журналах, то соответствующее уравнение прогнозирования действительно объясняет значимую долю отклонения в тарифах, на что указывает в результатах работы компьютерной программы p-значение 0,000 справа от значения F, равного 62,84.[6] Это говорит о том, что действительно обнаруживается устойчивая зависимость тарифов от этих факторов (или по крайней мере от одного из этих факторов), т.е. тарифы не являются чисто случайными величинами. Вам по-прежнему неизвестно, какие именно из этих X-переменных реально участвуют в прогнозировании Y, но вам доподлинно известно, что есть по крайней мере одна такая переменная.

Чтобы выяснить с помощью R^2, действительно ли уравнение регрессии является значимым, отметим, что коэффициент детерминации $R^2 = 0{,}787$, или 78,7%. Таблица R^2 для тестирования на уровне 5% в случае $n = 55$ журналов и $k = 3$ переменных (табл. 12.1.8) дает критическое значение 0,141, или 14,1%. Для того чтобы уравнение было значимым на привычном уровне 5%, X-переменные должны объяснять лишь 14,1% вариации тарифов (Y). Поскольку они объясняют больше, регрессию следует признать значимой.

[6] Когда в качестве p-значения указывается 0,000, его можно интерпретировать как $p < 0{,}0005$, поскольку p-значение, которое больше или равно 0,0005, будет округлено до 0,001.

Обратившись к таблицам R^2 для уровней 1% и 0,1% (табл. 12.1.9 и 12.1.10) при $n = 55$ и $k = 3$, находим критические значения 19,8% и 27,1% соответственно. Поскольку наблюдаемое значение коэффициента детерминации $R^2 = 78{,}7\%$ превосходит оба этих показателя, можно прийти к выводу, что эти X-переменные (величина читательской аудитории, процент мужчин и средний доход) имеют *очень высоко значимое влияние* на Y (тарифы). Используя терминологию p-значений, можно сказать, что регрессия в данном случае является очень высоко значимой ($p < 0{,}001$).

Чтобы убедиться в этом очень высоком уровне значимости, используя непосредственно F, можно сравнить F-статистику 62,84 (из компьютерной распечатки) со значением из F-таблицы для уровня 0,1% (табл. B.11 в приложении B), которое находится между 7,054 и 6,171 для $k = 3$ степеней свободы числителя и $n - k - 1 = 51$ степеней свободы знаменателя. (Поскольку значение 51 в таблице отсутствует, нам известно, что необходимое нам значение из F-таблицы находится в диапазоне от 7,054 для 30 степеней свободы знаменателя и 6,171 для 60 степеней свободы знаменателя.) Поскольку данная F-статистика (62,84) больше, чем значение из F-таблицы (значение из диапазона от 7,054 до 6,171), мы опять приходим к выводу, что полученный результат имеет очень высокую значимость ($p < 0{,}001$).

Таблицы критических значений для тестирования R^2

Таблицы 12.1.8–12.1.11 служат для тестирования значимости модели (F-тест). Эти таблицы позволяют проводить тестирование на уровнях 0,05 (значимый), 0,01(высоко значимый), 0,001 (очень высоко значимый) и 0,1. На каждом уровне тестирования регрессию можно считать значимой, если коэффициент детерминации R^2 превосходит значение из таблицы для имеющегося у вас количества X-переменных (k) и числа наблюдений (n). Если, например, вы имеете регрессию с $k = 2$ независимыми X-переменными и $n = 35$ наблюдениями, то она является значимой на уровне 0,05, при условии что R^2 превосходит критическое значение 0,171 (из таблицы для уровня 5%).

На практике большинство компьютерных программ автоматически выполняет F-тест и делает вывод относительно его значимости, а также, если тест значим, — об уровне значимости. В подобных случаях таблицы R^2 не нужны. Их использование преследует две цели: (1) выявить значимость, когда вы располагаете значением R^2, но у вас нет информации о результате проверке значимости, и (2) показать, насколько сильно уровень значимости зависит от n и k. Критическое значение R^2, на основе которого принимается решение о значимости, оказывается меньшим (менее "требовательным") при больших значениях n, поскольку в этом случае вы располагаете большей информацией. Однако критическое значение R^2, на основе которого принимается решение о значимости, оказывается большим (более "требовательным") при больших значениях k из-за усилий, необходимых для оценки дополнительных коэффициентов регрессии.

Если у вас более 60 наблюдений, критические значения можно найти с помощью двух множителей, указанных внизу таблицы R^2. Для этого необходимо воспользоваться следующей формулой.

Критические значения для R^2, когда $n > 60$

$$\text{Критическое значение} = \frac{\text{Множитель 1}}{n} + \frac{\text{Множитель 2}}{n^2}.$$

Таблица 12.1.8. Таблица R^2: критические значения для уровня 5% (значимо)

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	0,994									
4	0,902	0,997								
5	0,771	0,950	0,998							
6	0,658	0,864	0,966	0,999						
7	0,569	0,776	0,903	0,975	0,999					
8	0,499	0,698	0,832	0,924	0,980	0,999				
9	0,444	0,632	0,764	0,865	0,938	0,983	0,999			
10	0,399	0,575	0,704	0,806	0,887	0,947	0,985	0,999		
11	0,362	0,527	0,651	0,751	0,835	0,902	0,954	0,987	1,000	
12	0,332	0,486	0,604	0,702	0,785	0,856	0,914	0,959	0,989	1,000
13	0,306	0,451	0,563	0,657	0,739	0,811	0,872	0,924	0,964	0,990
14	0,283	0,420	0,527	0,618	0,697	0,768	0,831	0,885	0,931	0,967
15	0,264	0,393	0,495	0,582	0,659	0,729	0,791	0,847	0,896	0,937
16	0,247	0,369	0,466	0,550	0,624	0,692	0,754	0,810	0,860	0,904
17	0,232	0,348	0,440	0,521	0,593	0,659	0,719	0,775	0,825	0,871
18	0,219	0,329	0,417	0,494	0,564	0,628	0,687	0,742	0,792	0,839
19	0,208	0,312	0,397	0,471	0,538	0,600	0,657	0,711	0,761	0,807
20	0,197	0,297	0,378	0,449	0,514	0,574	0,630	0,682	0,731	0,777
21	0,187	0,283	0,361	0,429	0,492	0,550	0,604	0,655	0,703	0,749
22	0,179	0,270	0,345	0,411	0,471	0,527	0,580	0,630	0,677	0,722
23	0,171	0,259	0,331	0,394	0,452	0,507	0,558	0,607	0,653	0,696
24	0,164	0,248	0,317	0,379	0,435	0,488	0,538	0,585	0,630	0,673
25	0,157	0,238	0,305	0,364	0,419	0,470	0,518	0,564	0,608	0,650
26	0,151	0,229	0,294	0,351	0,404	0,454	0,501	0,545	0,588	0,629
27	0,145	0,221	0,283	0,339	0,390	0,438	0,484	0,527	0,569	0,609
28	0,140	0,213	0,273	0,327	0,377	0,424	0,468	0,510	0,551	0,590
29	0,135	0,206	0,264	0,316	0,365	0,410	0,453	0,495	0,534	0,573
30	0,130	0,199	0,256	0,306	0,353	0,397	0,439	0,480	0,518	0,556

Окончание табл. 12.1.8

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
31	0,126	0,193	0,248	0,297	0,342	0,385	0,426	0,466	0,503	0,540
32	0,122	0,187	0,240	0,288	0,332	0,374	0,414	0,452	0,489	0,525
33	0,118	0,181	0,233	0,279	0,323	0,363	0,402	0,440	0,476	0,511
34	0,115	0,176	0,226	0,271	0,314	0,353	0,391	0,428	0,463	0,497
35	0,111	0,171	0,220	0,264	0,305	0,344	0,381	0,417	0,451	0,484
36	0,108	0,166	0,214	0,257	0,297	0,335	0,371	0,406	0,440	0,472
37	0,105	0,162	0,208	0,250	0,289	0,326	0,362	0,396	0,429	0,461
38	0,103	0,157	0,203	0,244	0,282	0,318	0,353	0,386	0,418	0,449
39	0,100	0,153	0,198	0,238	0,275	0,310	0,344	0,377	0,408	0,439
40	0,097	0,150	0,193	0,232	0,268	0,303	0,336	0,368	0,399	0,429
41	0,095	0,146	0,188	0,226	0,262	0,296	0,328	0,359	0,390	0,419
42	0,093	0,142	0,184	0,221	0,256	0,289	0,321	0,351	0,381	0,410
43	0,090	0,139	0,180	0,216	0,250	0,283	0,314	0,344	0,373	0,401
44	0,088	0,136	0,176	0,211	0,245	0,276	0,307	0,336	0,365	0,393
45	0,086	0,133	0,172	0,207	0,239	0,271	0,300	0,329	0,357	0,384
46	0,085	0,130	0,168	0,202	0,234	0,265	0,294	0,322	0,350	0,377
47	0,083	0,127	0,164	0,198	0,230	0,259	0,288	0,316	0,343	0,369
48	0,081	0,125	0,161	0,194	0,225	0,254	0,282	0,310	0,336	0,362
49	0,079	0,122	0,158	0,190	0,220	0,249	0,277	0,304	0,330	0,355
50	0,078	0,120	0,155	0,186	0,216	0,244	0,272	0,298	0,323	0,348
51	0,076	0,117	0,152	0,183	0,212	0,240	0,267	0,293	0,318	0,342
52	0,075	0,115	0,149	0,180	0,208	0,235	0,262	0,287	0,312	0,336
53	0,073	0,113	0,146	0,176	0,204	0,231	0,257	0,282	0,306	0,330
54	0,072	0,111	0,143	0,173	0,201	0,227	0,252	0,277	0,301	0,324
55	0,071	0,109	0,141	0,170	0,197	0,223	0,248	0,272	0,295	0,318
56	0,069	0,107	0,138	0,167	0,194	0,219	0,244	0,267	0,290	0,313
57	0,068	0,105	0,136	0,164	0,190	0,215	0,240	0,263	0,285	0,308
58	0,067	0,103	0,134	0,161	0,187	0,212	0,236	0,258	0,281	0,303
59	0,066	0,101	0,131	0,159	0,184	0,208	0,232	0,254	0,276	0,298
60	0,065	0,100	0,129	0,156	0,181	0,205	0,228	0,250	0,272	0,293
Множитель 1	3,84	5,99	7,82	9,49	11,07	12,59	14,07	15,51	16,92	18,31
Множитель 2	2,15	−0,27	−3,84	−7,94	−12,84	−18,24	−23,78	−30,10	−36,87	−43,87

Таблица 12.1.9. Таблица R^2: критические значения для уровня 1% (высоко значимо)

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	1,000									
4	0,980	1,000								
5	0,919	0,990	1,000							
6	0,841	0,954	0,993	1,000						
7	0,765	0,900	0,967	0,995	1,000					
8	0,696	0,842	0,926	0,975	0,996	1,000				
9	0,636	0,785	0,879	0,941	0,979	0,997	1,000			
10	0,585	0,732	0,830	0,901	0,951	0,982	0,997	1,000		
11	0,540	0,684	0,784	0,859	0,916	0,958	0,985	0,997	1,000	
12	0,501	0,641	0,740	0,818	0,879	0,928	0,963	0,987	0,998	1,000
13	0,467	0,602	0,700	0,778	0,842	0,894	0,936	0,967	0,988	0,998
14	0,437	0,567	0,663	0,741	0,806	0,860	0,906	0,943	0,971	0,989
15	0,411	0,536	0,629	0,706	0,771	0,827	0,875	0,915	0,948	0,973
16	0,388	0,508	0,598	0,673	0,738	0,795	0,844	0,887	0,923	0,953
17	0,367	0,482	0,570	0,643	0,707	0,764	0,814	0,858	0,896	0,929
18	0,348	0,459	0,544	0,616	0,678	0,734	0,784	0,829	0,869	0,904
19	0,331	0,438	0,520	0,590	0,652	0,707	0,757	0,802	0,843	0,879
20	0,315	0,418	0,498	0,566	0,626	0,681	0,730	0,775	0,816	0,854
21	0,301	0,401	0,478	0,544	0,603	0,656	0,705	0,750	0,791	0,829
22	0,288	0,384	0,459	0,523	0,581	0,633	0,681	0,726	0,767	0,805
23	0,276	0,369	0,442	0,504	0,560	0,612	0,659	0,703	0,744	0,782
24	0,265	0,355	0,426	0,487	0,541	0,591	0,638	0,681	0,721	0,759
25	0,255	0,342	0,410	0,470	0,523	0,572	0,618	0,660	0,700	0,738
26	0,246	0,330	0,396	0,454	0,506	0,554	0,599	0,641	0,680	0,717
27	0,237	0,319	0,383	0,440	0,490	0,537	0,581	0,622	0,661	0,698
28	0,229	0,308	0,371	0,426	0,475	0,521	0,564	0,605	0,643	0,679
29	0,221	0,298	0,359	0,413	0,461	0,506	0,548	0,588	0,625	0,661
30	0,214	0,289	0,349	0,401	0,448	0,492	0,533	0,572	0,609	0,644
31	0,208	0,280	0,338	0,389	0,435	0,478	0,519	0,557	0,593	0,627
32	0,201	0,272	0,329	0,378	0,423	0,465	0,505	0,542	0,578	0,612
33	0,195	0,264	0,319	0,368	0,412	0,453	0,492	0,529	0,563	0,597
34	0,190	0,257	0,311	0,358	0,401	0,442	0,479	0,515	0,550	0,583

Окончание табл. 12.1.9

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
35	0,185	0,250	0,303	0,349	0,391	0,430	0,468	0,503	0,537	0,569
36	0,180	0,244	0,295	0,340	0,381	0,420	0,456	0,491	0,524	0,556
37	0,175	0,237	0,287	0,332	0,372	0,410	0,446	0,480	0,512	0,543
38	0,170	0,231	0,280	0,324	0,363	0,400	0,435	0,469	0,501	0,531
39	0,166	0,226	0,274	0,316	0,355	0,391	0,426	0,458	0,490	0,520
40	0,162	0,220	0,267	0,309	0,347	0,382	0,416	0,448	0,479	0,509
41	0,158	0,215	0,261	0,302	0,339	0,374	0,407	0,439	0,469	0,498
42	0,155	0,210	0,255	0,295	0,332	0,366	0,399	0,430	0,459	0,488
43	0,151	0,206	0,250	0,289	0,325	0,358	0,390	0,421	0,450	0,478
44	0,148	0,201	0,244	0,283	0,318	0,351	0,382	0,412	0,441	0,469
45	0,145	0,197	0,239	0,277	0,311	0,344	0,375	0,404	0,432	0,460
46	0,141	0,193	0,234	0,271	0,305	0,337	0,367	0,396	0,424	0,451
47	0,138	0,189	0,230	0,266	0,299	0,330	0,360	0,389	0,416	0,443
48	0,136	0,185	0,225	0,261	0,293	0,324	0,353	0,381	0,408	0,435
49	0,133	0,181	0,221	0,256	0,288	0,318	0,347	0,374	0,401	0,427
50	0,130	0,178	0,217	0,251	0,283	0,312	0,341	0,368	0,394	0,419
51	0,128	0,175	0,213	0,246	0,278	0,307	0,335	0,361	0,387	0,412
52	0,125	0,171	0,209	0,242	0,273	0,301	0,329	0,355	0,381	0,405
53	0,123	0,168	0,205	0,238	0,268	0,296	0,323	0,349	0,374	0,398
54	0,121	0,165	0,201	0,233	0,263	0,291	0,318	0,343	0,368	0,391
55	0,119	0,162	0,198	0,229	0,259	0,286	0,312	0,337	0,362	0,385
56	0,117	0,160	0,194	0,226	0,254	0,281	0,307	0,332	0,356	0,379
57	0,115	0,157	0,191	0,222	0,250	0,277	0,302	0,326	0,350	0,373
58	0,113	0,154	0,188	0,218	0,246	0,272	0,297	0,321	0,345	0,367
59	0,111	0,152	0,185	0,215	0,242	0,268	0,293	0,316	0,339	0,361
60	0,109	0,149	0,182	0,211	0,238	0,264	0,288	0,311	0,334	0,356
Множитель 1	6,63	9,21	11,35	13,28	15,09	16,81	18,48	20,09	21,67	23,21
Множитель 2	−5,81	−15,49	−25,66	−36,39	−47,63	−59,53	−71,65	−84,60	−97,88	−111,76

Таблица 12.1.10. Таблица R^2: критические значения для уровня 0,1% (в высшей степени значимо)

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	1,000									
4	0,998	1,000								
5	0,982	0,999	1,000							
6	0,949	0,990	0,999	1,000						
7	0,904	0,968	0,993	0,999	1,000					
8	0,855	0,937	0,977	0,995	1,000	1,000				
9	0,807	0,900	0,952	0,982	0,996	1,000	1,000			
10	0,761	0,861	0,922	0,961	0,985	0,996	1,000	1,000		
11	0,717	0,822	0,889	0,936	0,967	0,987	0,997	1,000	1,000	
12	0,678	0,785	0,856	0,908	0,945	0,972	0,989	0,997	1,000	1,000
13	0,642	0,749	0,822	0,878	0,920	0,952	0,975	0,990	0,997	1,000
14	0,608	0,715	0,790	0,848	0,894	0,930	0,958	0,978	0,991	0,998
15	0,578	0,684	0,759	0,819	0,867	0,906	0,938	0,962	0,980	0,992
16	0,550	0,654	0,730	0,790	0,840	0,881	0,916	0,944	0,966	0,982
17	0,525	0,627	0,702	0,763	0,813	0,856	0,893	0,923	0,949	0,968
18	0,502	0,602	0,676	0,736	0,787	0,831	0,869	0,902	0,930	0,953
19	0,480	0,578	0,651	0,711	0,763	0,807	0,846	0,880	0,910	0,935
20	0,461	0,556	0,628	0,688	0,739	0,784	0,824	0,859	0,890	0,917
21	0,442	0,536	0,606	0,665	0,716	0,761	0,801	0,837	0,869	0,897
22	0,426	0,517	0,586	0,644	0,694	0,739	0,780	0,816	0,849	0,878
23	0,410	0,499	0,567	0,624	0,674	0,718	0,759	0,795	0,829	0,859
24	0,395	0,482	0,548	0,605	0,654	0,698	0,739	0,775	0,809	0,839
25	0,382	0,466	0,531	0,587	0,635	0,679	0,719	0,756	0,790	0,821
26	0,369	0,452	0,515	0,570	0,618	0,661	0,701	0,737	0,771	0,802
27	0,357	0,438	0,500	0,553	0,601	0,644	0,683	0,719	0,753	0,784
28	0,346	0,425	0,486	0,538	0,585	0,627	0,666	0,702	0,735	0,767
29	0,335	0,412	0,472	0,523	0,569	0,611	0,649	0,685	0,718	0,750
30	0,325	0,401	0,459	0,510	0,555	0,596	0,634	0,669	0,702	0,733
31	0,316	0,389	0,447	0,496	0,541	0,581	0,619	0,654	0,686	0,717

Окончание табл. 12.1.10

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
32	0,307	0,379	0,435	0,484	0,527	0,567	0,604	0,639	0,671	0,702
33	0,299	0,369	0,424	0,472	0,515	0,554	0,590	0,625	0,657	0,687
34	0,291	0,360	0,414	0,460	0,503	0,541	0,577	0,611	0,643	0,673
35	0,283	0,351	0,404	0,450	0,491	0,529	0,564	0,598	0,629	0,659
36	0,276	0,342	0,394	0,439	0,480	0,517	0,552	0,585	0,616	0,646
37	0,269	0,334	0,385	0,429	0,469	0,506	0,540	0,573	0,604	0,633
38	0,263	0,326	0,376	0,420	0,459	0,495	0,529	0,561	0,591	0,620
39	0,257	0,319	0,368	0,411	0,449	0,485	0,518	0,550	0,580	0,608
40	0,251	0,312	0,360	0,402	0,440	0,475	0,508	0,539	0,569	0,597
41	0,245	0,305	0,352	0,393	0,431	0,465	0,498	0,529	0,558	0,586
42	0,240	0,298	0,345	0,385	0,422	0,456	0,488	0,518	0,547	0,575
43	0,235	0,292	0,338	0,378	0,414	0,447	0,479	0,509	0,537	0,564
44	0,230	0,286	0,331	0,370	0,406	0,439	0,470	0,499	0,527	0,554
45	0,225	0,280	0,324	0,363	0,398	0,431	0,461	0,490	0,518	0,544
46	0,220	0,275	0,318	0,356	0,391	0,423	0,453	0,482	0,509	0,535
47	0,216	0,269	0,312	0,349	0,383	0,415	0,445	0,473	0,500	0,526
48	0,212	0,264	0,306	0,343	0,377	0,408	0,437	0,465	0,491	0,517
49	0,208	0,259	0,301	0,337	0,370	0,401	0,429	0,457	0,483	0,508
50	0,204	0,255	0,295	0,331	0,363	0,394	0,422	0,449	0,475	0,500
51	0,200	0,250	0,290	0,325	0,357	0,387	0,415	0,442	0,467	0,492
52	0,197	0,246	0,285	0,320	0,351	0,381	0,408	0,435	0,460	0,484
53	0,193	0,242	0,280	0,314	0,345	0,374	0,402	0,428	0,453	0,477
54	0,190	0,237	0,276	0,309	0,340	0,368	0,395	0,421	0,446	0,469
55	0,186	0,233	0,271	0,304	0,334	0,362	0,389	0,414	0,439	0,462
56	0,183	0,230	0,267	0,299	0,329	0,357	0,383	0,408	0,432	0,455
57	0,180	0,226	0,262	0,294	0,324	0,351	0,377	0,402	0,426	0,448
58	0,177	0,222	0,258	0,290	0,319	0,346	0,371	0,396	0,419	0,442
59	0,174	0,219	0,254	0,285	0,314	0,341	0,366	0,390	0,413	0,436
60	0,172	0,215	0,250	0,281	0,309	0,336	0,361	0,384	0,407	0,429
Множитель 1	10,83	13,82	16,27	18,47	20,52	22,46	24,32	26,12	27,88	29,59
Множитель 2	−31,57	−54,02	−75,12	−96,26	−117,47	−138,94	−160,86	−183,33	−206,28	−229,55

Таблица 12.1.11. Таблица R^2: критические значения для уровня 10%

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	0,976									
4	0,810	0,990								
5	0,649	0,900	0,994							
6	0,532	0,785	0,932	0,996						
7	0,448	0,684	0,844	0,949	0,997					
8	0,386	0,602	0,759	0,877	0,959	0,997				
9	0,339	0,536	0,685	0,804	0,898	0,965	0,998			
10	0,302	0,482	0,622	0,738	0,835	0,914	0,970	0,998		
11	0,272	0,438	0,568	0,680	0,775	0,857	0,925	0,974	0,998	
12	0,247	0,401	0,523	0,628	0,721	0,803	0,874	0,933	0,977	0,998
13	0,227	0,369	0,484	0,584	0,673	0,753	0,825	0,888	0,940	0,979
14	0,209	0,342	0,450	0,545	0,630	0,708	0,779	0,842	0,899	0,946
15	0,194	0,319	0,420	0,510	0,592	0,667	0,736	0,799	0,857	0,907
16	0,181	0,298	0,394	0,480	0,558	0,630	0,697	0,759	0,816	0,868
17	0,170	0,280	0,371	0,453	0,527	0,596	0,661	0,721	0,778	0,830
18	0,160	0,264	0,351	0,428	0,499	0,566	0,628	0,687	0,742	0,794
19	0,151	0,250	0,332	0,406	0,474	0,538	0,598	0,655	0,709	0,760
20	0,143	0,237	0,316	0,386	0,452	0,513	0,571	0,626	0,679	0,729
21	0,136	0,226	0,301	0,368	0,431	0,490	0,546	0,599	0,650	0,699
22	0,129	0,215	0,287	0,352	0,412	0,469	0,523	0,575	0,624	0,671
23	0,124	0,206	0,275	0,337	0,395	0,450	0,502	0,552	0,600	0,646
24	0,118	0,197	0,263	0,323	0,379	0,432	0,482	0,530	0,577	0,622
25	0,113	0,189	0,253	0,310	0,364	0,415	0,464	0,511	0,556	0,599
26	0,109	0,181	0,243	0,298	0,350	0,400	0,447	0,492	0,536	0,579
27	0,105	0,175	0,234	0,287	0,338	0,386	0,431	0,475	0,518	0,559
28	0,101	0,168	0,225	0,277	0,326	0,372	0,417	0,459	0,501	0,541
29	0,097	0,162	0,218	0,268	0,315	0,360	0,403	0,444	0,484	0,523
30	0,094	0,157	0,210	0,259	0,305	0,348	0,390	0,430	0,469	0,507
31	0,091	0,152	0,203	0,251	0,295	0,337	0,378	0,417	0,455	0,492
32	0,088	0,147	0,197	0,243	0,286	0,327	0,366	0,405	0,442	0,478
33	0,085	0,142	0,191	0,236	0,277	0,317	0,356	0,393	0,429	0,464
34	0,082	0,138	0,185	0,229	0,269	0,308	0,346	0,382	0,417	0,451

Окончание табл. 12.1.11

Количество наблюдений (n)	Количество X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
35	0,080	0,134	0,180	0,222	0,262	0,300	0,336	0,371	0,406	0,439
36	0,078	0,130	0,175	0,216	0,255	0,291	0,327	0,361	0,395	0,427
37	0,075	0,127	0,170	0,210	0,248	0,284	0,318	0,352	0,385	0,416
38	0,073	0,123	0,166	0,205	0,241	0,276	0,310	0,343	0,375	0,406
39	0,071	0,120	0,162	0,199	0,235	0,269	0,302	0,334	0,366	0,396
40	0,070	0,117	0,157	0,194	0,229	0,263	0,295	0,326	0,357	0,387
41	0,068	0,114	0,154	0,190	0,224	0,257	0,288	0,319	0,348	0,378
42	0,066	0,111	0,150	0,185	0,219	0,250	0,281	0,311	0,340	0,369
43	0,065	0,109	0,146	0,181	0,214	0,245	0,275	0,304	0,333	0,361
44	0,063	0,106	0,143	0,177	0,209	0,239	0,269	0,297	0,325	0,353
45	0,062	0,104	0,140	0,173	0,204	0,234	0,263	0,291	0,318	0,345
46	0,060	0,102	0,137	0,169	0,200	0,229	0,257	0,285	0,312	0,338
47	0,059	0,099	0,134	0,166	0,196	0,224	0,252	0,279	0,305	0,331
48	0,058	0,097	0,131	0,162	0,191	0,220	0,247	0,273	0,299	0,324
49	0,057	0,095	0,128	0,159	0,188	0,215	0,242	0,268	0,293	0,318
50	0,055	0,093	0,126	0,156	0,184	0,211	0,237	0,263	0,287	0,312
51	0,054	0,092	0,123	0,153	0,180	0,207	0,233	0,258	0,282	0,306
52	0,053	0,090	0,121	0,150	0,177	0,203	0,228	0,253	0,277	0,300
53	0,052	0,088	0,119	0,147	0,174	0,199	0,224	0,248	0,272	0,295
54	0,051	0,086	0,116	0,144	0,170	0,196	0,220	0,244	0,267	0,290
55	0,050	0,085	0,114	0,142	0,167	0,192	0,216	0,239	0,262	0,284
56	0,049	0,083	0,112	0,139	0,164	0,189	0,212	0,235	0,257	0,279
57	0,049	0,082	0,110	0,137	0,162	0,185	0,209	0,231	0,253	0,275
58	0,048	0,080	0,108	0,134	0,159	0,182	0,205	0,227	0,249	0,270
59	0,047	0,079	0,107	0,132	0,156	0,179	0,202	0,223	0,245	0,266
60	0,046	0,078	0,105	0,130	0,153	0,176	0,198	0,220	0,241	0,261
Множитель 1	2,71	4,61	6,25	7,78	9,24	10,65	12,02	13,36	14,68	15,99
Множитель 2	3,12	3,08	2,00	0,32	–1,92	–4,75	–7,59	–11,12	–14,94	–19,05

Например, при $n = 135$ наблюдений и $k = 6$, объясняющих X-переменных, чтобы выполнить тестирование на уровне 0,05, нужно использовать два множителя 12,59 и –18,24 внизу столбца для $k = 6$ таблицы для уровня 5%. Воспользовавшись приведенной выше формулой, найдем соответствующее критическое значение для R^2:

$$\text{Критическое значение} = \frac{\text{Множитель 1}}{n} + \frac{\text{Множитель 2}}{n^2} =$$

$$= \frac{12,59}{135} + \frac{-18,24}{135^2} = 0,09326 - 0,00100 = 0,0923.$$

Если R^2 для вашей совокупности данных (из компьютерной распечатки) превосходит это значение (0,0923, или 9,23%), F-тест является значимым; в противном случае — нет.

Какие переменные являются значимыми: t-тест для каждого коэффициента

Если F-тест является значимым, то вам известно, что одна или несколько X-переменных могут быть полезны в прогнозировании Y и, следовательно, можно продолжать анализ с помощью t-тестов для отдельных коэффициентов регрессии с целью выяснять, какие именно из X-переменных действительно полезны. Эти t-тесты определяют, оказывает ли значимое влияние на Y та или иная X-переменная, если все другие X-переменные остаются при этом неизменными. Следует помнить, что, приняв нулевую гипотезу, вы сделали *слабое* заключение и, по сути, тем самым не доказали бесполезность X-переменной, а просто у вас не хватило убедительных доказательств наличия взаимосвязи. Таким образом, взаимосвязь может существовать, но вследствие действия фактора случайности или из-за небольшого размера выборки вы не в состоянии обнаружить ее с помощью тех данных, которые имеются в вашем распоряжении.

Если же F-тест не является значимым, то использовать t-тесты для отдельных коэффициентов регрессии нельзя. В редких случаях эти t-тесты могут быть значимыми даже тогда, когда F-тест не является значимым. При этом F-тест считается более важным и необходимо делать вывод о том, что все коэффициенты являются незначимыми. Поступив иначе, вы повысите ошибку I рода выше объявленного уровня (например, 5%).

t-тест для каждого коэффициента основан на оценке коэффициента регрессии и его стандартной ошибке и использует критическое значение из t-таблицы для $n - k - 1$ степеней свободы. Доверительный интервал для какого-либо конкретного коэффициента регрессии в генеральной совокупности (например, j-го — β_j) определяется обычным способом.

Доверительный интервал для j-го коэффициента регрессии, β_j

$$\text{От } b_j - tS_{b_j} \text{ до } b_j + tS_{b_j},$$

где t берется из t-таблицы для $n - k - 1$ степеней свободы.

t-тест является значимым, если заданное значение 0 (указывающее на отсутствие влияния) *не* попадает в этот доверительный интервал. Здесь нет ничего нового: это обычная процедура для двустороннего тестирования.

Как альтернативный вариант можно сравнить t-статистику b_j / S_{b_j} со значением из t-таблицы и сделать вывод о значимости, если абсолютное значение этой t-статистики оказывается больше. Если посмотреть на последние значения в каж-

дом из столбцов t-таблицы, можно увидеть достаточно простой, приблизительный способ определения значимости коэффициентов: значимыми будут те коэффициенты регрессии, для которых t-статистика по абсолютному значению равна или больше 2, поскольку для достаточно больших n и уровня значимости 5% значение из t-таблицы приблизительно равно 2. Как всегда, оба метода, и на использовании t-статистики, и на использовании доверительного интервала, должны в любом случае обеспечивать одинаковый результат (значимость или не значимость) для каждого теста.

Что же именно в данном случае тестируется? В результате t-теста для β_j мы должны принять решение, оказывает ли X_j значимое влияние на Y в исследуемой генеральной совокупности, когда все другие X-переменные остаются неизменными. В этом случае речь не идет о корреляции между X_j и Y, которая игнорирует все остальные X-переменные. Скорее, это проверка влияния X_j на Y после внесения поправки на все остальные факторы. Например, в исследованиях уровня заработной платы, цель которых заключается в выявлении возможных фактов дискриминации по признаку пола, обычно делают поправку на уровень образования и стаж работы. Несмотря на то что мужчины в компании могут (в среднем) получать более высокую заработную плату, чем женщины, очень важно понять, не объясняются ли эти различия какими-либо другими факторами, помимо пола. В результате включения всех этих факторов в множественную регрессию (регрессия Y = заработная плата на X_1 = пол, X_2 = образование и X_3 = стаж работы) коэффициент регрессии для пола будет отражать влияние пола на уровень заработной платы с учетом поправок на уровень образования и стаж работы.[7]

Ниже приведены формулы для гипотез, касающихся проверки значимости j-го коэффициента регрессии.

Гипотезы для t-теста j-го коэффициента регрессии

$$H_0: \beta_j = 0;$$
$$H_1: \beta_j \neq 0.$$

Если вернуться к нашему примеру с тарифами на размещение рекламных объявлений в журналах, то соответствующий t-тест будет иметь $n - k - 1 = 55 - 3 - 1 = 51$ степеней свободы. Двустороннее критическое значение из t-таблицы равно 1,960 (или, точнее, 2,008).[8] В табл. 12.1.12 приведена соответствующая информация из компьютерной распечатки в табл. 12.1.6.

[7] Переменную пола, X_1, можно представить как 0 — для женщин и 1 — для мужчин. В таком случае коэффициент регрессии будет представлять дополнительную оплату в среднем для мужчины в сравнении с женщиной, имеющей тот же уровень образования и стаж работы. Если же переменную пола представить как 1 — для женщин и 0 — для мужчин, то коэффициент регрессии будет представлять дополнительную оплату для женщины в сравнении с мужчиной, имеющим тот же уровень образования и стаж работы. К счастью, выводы окажутся одинаковыми, независимо от того, каким представлением мы будем пользоваться.

[8] Помните, что использование t-значения для бесконечного числа степеней свободы (т.е. в случае, когда речь идет о 40 и более степенях свободы) представляет собой лишь аппроксимацию. В этом случае истинное значение из t-таблицы равно 2,008, а 1,960 — лишь удобное приближение.

Две из трех X-переменных являются значимыми, поскольку для них p-значения оказываются меньше 0,05. Еще один (эквивалентный) способ проверки значимости заключается в том, чтобы выяснить, какие t-статистики (в компьютерной распечатке соответствующий столбец обозначен просто t) оказываются большими, чем 2,008. И еще один (тоже эквивалентный) способ проверки значимости состоит в том, чтобы выяснить, какие из 95% доверительных интервалов для коэффициентов регрессии не включают 0. Как мы и предполагали ранее, величина читательской аудитории оказывает огромное влияние на рекламные тарифы в журналах. Столь высокое значение t (13,48) означает, что влияние величины читательской аудитории на рекламные тарифы является очень высоко значимым (при условии, что процент читателей-мужчин и средний доход остаются постоянными). Влияние среднего дохода на рекламные тарифы в журналах также является значимым (при условии, что процент читателей-мужчин и величина читательской аудитории остаются постоянными).

Очевидно, что процент читателей-мужчин не оказывает на тарифы значительного влияния (при условии, что величина читательской аудитории и средний доход остаются постоянными), поскольку соответствующий t-тест не является значимым. Не исключено, что этот процент оказывает на тарифы определенное влияние только через доход (средний доход у мужчин может быть выше, чем у женщин). Таким образом, после внесения поправки на средний доход можно ожидать, что переменная, соответствующая проценту мужчин, уже не будет нести дополнительной информации для прогнозирования тарифов. Несмотря на то что оцениваемое влияние процента читателей-мужчин составляет −$123,6, его отклонение от 0 носит лишь случайный характер. Строго говоря, этот коэффициент, −$123,6, не подлежит интерпретации; поскольку он не является значимым, вы "не имеете права" объяснять его. Иными словами, его значение (−$123,6) — лишь видимость, и, по сути, ничем не отличается от $0,00; более того, в действительности вы не можете даже сказать, положительное это число или отрицательное!

Константа, a = $4 043, не является значимой. Она не отличается существенно от нуля. Нельзя сказать ничего определенного и о знаке соответствующего параметра генеральной совокупности, a, поскольку его вполне можно считать равным нулю. В приложениях, связанных с калькуляцией затрат, a зачастую служит оценкой фиксированных затрат производства. Доверительные интервалы и проверки гипотез покажут вам, существует ли в действительности значимый фиксированный компонент в вашей структуре затрат.

Таблица 12.1.12. Компьютерная распечатка результатов множественной регрессии

Независимая переменная	Коэффициент	Стандартное отклонение	t	p
Константа	4043	16884	0,24	0,812
Аудитория	3,7880	0,2809	13,48	0,000
Процент мужчин	−123,6	137,8	−0,90	0,374
Доход	0,9026	0,3696	2,44	0,018

Другие проверки, касающиеся коэффициента регрессии

Другие проверки применительно к коэффициенту регрессии можно выполнить точно так же, как это делается и в случае средних значений. Если для одного из коэффициентов регрессии существует некоторое заданное значение (источником которого не являются рассматриваемые данные), можно проверить, значимо ли оценка коэффициента регрессии отличается от этого заданного значения. Для этого достаточно проверить, попадает ли это заданное значение в доверительный интервал, и если не попадает, то, как обычно, принять решение о "значимом отличии". В качестве альтернативного варианта можно воспользоваться t-статистикой (b_j – заданное значение)/ S_{b_j}, приняв решение о "значимом отличии", если абсолютное значение этой статистики превышает значение из t-таблицы для $n - k - 1$ степеней свободы.

Допустим, вы решили (до того как вам встретилась эта совокупность данных), что дополнительные затраты на рекламу составляют \$2,00 из расчета на каждую тысячу человек. Чтобы проверить это предположение, можно использовать \$2,00 в качестве заданного значения. Поскольку доверительный интервал расходов на рекламу для читательской аудитории (от \$3,22 до \$4,35) не включает это заданное значение, можно прийти к выводу, что влияние читательской аудитории на рекламные затраты (с поправкой на процент читателей-мужчин и средний доход) оказывается существенно большим, чем \$2,00 на каждую тысячу человек. Обратите внимание, что мы делаем односторонний вывод на основе двустороннего теста. Двусторонний тест в данном случае вполне уместен, поскольку соответствующая оценка могла бы оказаться и по другую сторону от значения \$2,00.

Односторонние доверительные интервалы можно вычислять обычным способом для одного (или нескольких) коэффициента регрессии, что дает возможность делать одностороннее утверждение об интересующем нас коэффициенте (или о коэффициентах) регрессии для соответствующей генеральной совокупности. При этом следует обязательно использовать односторонние t-значения из t-таблицы для $n - k - 1$ степеней свободы, а доверительный интервал для β_j обязательно должен включать коэффициент регрессии b_j.

Например, коэффициент регрессии для дохода равняется $b_3 = 0{,}9026$, а это свидетельствует о том, что (при всех прочих равных условиях) каждый дополнительный доллар среднего дохода приводит к повышению цены полностраничного рекламного объявления в среднем на \$0,9026. Стандартная ошибка равно $S_{b_3} = 0{,}3696$, а одностороннее значение из t-таблицы для $n - k - 1 = 51$ степеней свободы составляет $t = 1{,}645$, поэтому нижней границей одностороннего интервала является $b_3 - tS_{b_3} = 0{,}9026 - 1{,}645 \times 0{,}3696 = 0{,}2946$. Ваше заключение будет иметь следующий вид.

"Мы на 95% уверены в том, что каждый дополнительный доллар среднего дохода приводит к повышению средних затрат на страницу рекламы по меньшей мере на \$0,29".

Эти 29 центов, определяющие одностороннюю доверительную границу, оказываются намного меньше, чем оценочное значение \$0,9026, поскольку мы сделали поправку на случайные ошибки оценки. Воспользовавшись не двусторонним, а односторонним интервалом, мы можем принять решение о 29, а не о 16 центах, которые определяют нижнюю границу двустороннего интервала.

Односторонние тесты по отношению к коэффициентам регрессии можно выполнять обычным способом при условии, что вас интересует лишь *одна сторона* эталонного значения и что вы не измените эту интересующую вас сторону, если оценки окажутся другими.

Какие переменные оказывают большее влияние

Какая из X-переменных оказывает наибольшее влияние на Y? Хороший вопрос! К сожалению, исчерпывающего ответа на этот вопрос нет, ввиду того, что наличие взаимосвязей между X-переменными может сделать принципиально невозможным выяснение того, какая именно из X-переменных в действительности "отвечает" за поведение переменной Y. Ответ на поставленный вопрос зависит от конкретной ситуации (в частности, можно ли изменять X-переменные по отдельности). Ответ определяется также наличием взаимосвязи (или корреляции) между X-переменными. Ниже мы рассмотрим два полезных (хоть и неполных) ответа на этот непростой вопрос.

Сравнение стандартизованных коэффициентов регрессии

Поскольку все коэффициенты регрессии $b_1, \ldots, b_k$ могут быть выражены в разных единицах измерения, непосредственное их сравнение весьма затруднительно: небольшой коэффициент может на самом деле оказаться более важным, чем большой. Короче говоря, здесь мы имеем дело с классической проблемой "попытки сравнения яблок и апельсинов". *Стандартизованные коэффициенты регрессии* позволяют решить эту проблему за счет представления коэффициентов регрессии в терминах единого множества имеющих статистический смысл единиц измерения, что позволяет по крайней мере попытаться проводить сравнение.

Коэффициент регрессии b_i указывает влияние изменения X_i на переменную Y, когда все другие X-переменные остаются неизменными. Коэффициент регрессии b_i измеряется в единицах измерения Y на одну единицу измерения X_i. Если, например, Y представляет собой объем продаж в долларовом выражении, а X_1 — количество торгового персонала, то b_1 выражается в количестве долларов (объем продаж) на одного человека. Допустим, что следующий коэффициент регрессии, b_2, выражается в количестве долларов (объем продаж) на суммарный километраж рабочих поездок торговых представителей компании. Непосредственное сравнение b_1 и b_2 не позволит нам ответить на вопрос, какой из этих двух факторов (уровень торгового персонала или командировочные расходы компании) оказывает большее влияние на объем продаж, потому что разные единицы измерения (доллары на человека и доллары на километр) непосредственно сравнивать нельзя.

Стандартизованный коэффициент регрессии, который вычисляется путем умножения коэффициента регрессии b_i на S_{X_i} и деления полученного произведения на S_Y, представляет собой ожидаемое изменение Y (в стандартизованных единицах S_Y), вызванное увеличением X_i на одну соответствующую стандартизованную единицу (т.е. S_{X_i}), когда все другие X-переменные остаются неизменны-

ми.[9] Абсолютные значения стандартизованных коэффициентов регрессии можно сравнивать, получая при этом некоторое представление об относительной важности соответствующих переменных.[10] Каждый стандартизованный коэффициент регрессии измеряется в единицах стандартных отклонений Y на одно стандартное отклонение X_i. Это обычные выборочные стандартные отклонения для каждой переменной, о которых мы уже говорили в главе 5. Использование таких единиц вполне естественно, поскольку они создают шкалу измерений, соответствующую фактической вариации каждой переменной в вашей совокупности данных.

Стандартизованный коэффициент регрессии

$$b_i S_{X_i} / S_Y .$$

Каждый коэффициент регрессии корректируется с помощью отношения обычных выборочных стандартных отклонений. Абсолютные значения позволяют получить приблизительное представление об относительной важности X-переменных.

Чтобы стандартизировать коэффициенты регрессии в примере о рекламных объявлений в журналах, нужно сначала вычислить стандартные отклонения для каждой из переменных, как показано ниже.

Стандартные отклонения			
Стоимость страницы	Читательская аудитория	Процент читателей-мужчин	Средний доход
$S_Y = 45446$	$S_{X_1} = 11212$	$S_{X_2} = 25{,}883$	$S_{X_3} = 10225$

Вам также требуются коэффициенты регрессии, приведенные ниже.

Коэффициенты регрессии		
Читательская аудитория	Процент читателей-мужчин	Средний доход
$b_1 = 3{,}7880$	$b_2 = -123{,}6$	$b_3 = 0{,}9026$

Наконец, можно вычислить стандартизованные коэффициенты регрессии.

Стандартизованные коэффициенты регрессии		
Читательская аудитория	Процент читателей-мужчин	Средний доход
$b_1 S_{X_1} / S_Y =$	$b_2 S_{X_2} / S_Y =$	$b_3 S_{X_3} / S_Y =$
$= 3{,}7880 \times 11212 / 45446 =$	$= -123{,}6 \times 25{,}883 / 45446 =$	$= 0{,}9026 \times 10225 / 45446 =$
$= 0{,}935$	$= -0{,}070$	$= 0{,}203$

[9] Стандартизованные коэффициенты регрессии иногда называют *бета-коэффициентами*. Мы постараемся не прибегать к использованию этого термина, поскольку его легко спутать с коэффициентами регрессии в генеральной совокупности (также β, или *бета*) и недиверсифицируемым компонентом риска в финансах (который называется *бета* ценных бумаг и представляет собой обычный, нестандартизованный коэффициент регрессии выборки, где X является процентным изменением рыночного индекса, а Y — процентным изменением стоимости сертификата на ценные бумаги).

[10] Напомним, что абсолютное значение просто игнорирует знак "минус".

Приведем непосредственную интерпретацию одного из этих стандартизованных коэффициентов. Значение 0,935, относящееся к читательской аудитории, свидетельствует о том, что увеличение аудитории на одно ее стандартное отклонение (11 212) приведет к ожидаемому увеличению тарифа на размещение рекламы в журналах на 0,935 его (тарифа) стандартных отклонений (45 446). Иными словами, увеличение аудитории на 11 212 (одно стандартное отклонение) приведет к ожидаемому увеличению тарифа на размещение рекламы в журналах на $0{,}935 \times 45\,446 = \$42\,492$ (несколько меньше, 0,935, чем одно стандартное отклонение тарифа на размещение рекламы).

Гораздо важнее, однако, то обстоятельство, что эти стандартизованные коэффициенты регрессии теперь можно сравнивать между собой. Наибольшим по абсолютному значению является коэффициент 0,935 для читательской аудитории; это свидетельствует о том, что данная переменная является самой важной из трех X-переменных. Далее следует средний доход, для которого коэффициент равен 0,203. Наименьшее абсолютное значение коэффициента $|-0{,}070| = 0{,}070$ соответствует проценту читателей-мужчин.

Было бы неправильным сравнивать коэффициенты регрессии непосредственно, не стандартизовав их предварительно. Обратите внимание, что проценту читателей-мужчин соответствует наибольший (по абсолютному значению) коэффициент регрессии, $|-123{,}6| = 123{,}6$. Однако поскольку он выражается в единицах измерения, отличных от единиц измерения других коэффициентов регрессии, непосредственное сравнение лишено смысла.

Абсолютные значения стандартизованных коэффициентов регрессии можно сравнивать друг с другом, что позволяет получить *грубое* представление о важности соответствующих переменных. Еще раз следует подчеркнуть, что эти результаты не являются идеальными, поскольку взаимосвязи между X-переменными могут сделать принципиально невозможным выяснение того, какая из X-переменных в действительности "отвечает" за поведение переменной Y.

Сравнение коэффициентов корреляции

Нас вообще могут не очень-то интересовать коэффициенты регрессии, полученные из множественной регрессии и представляющие влияние каждой переменной при условии, что все другие переменные остаются неизменными. Если нас интересует лишь то, в какой мере каждая из X-переменных влияет на Y при условии, что все другие X-переменные продолжают "вести себя естественным образом" (т.е. мы *не* пытаемся принудительно зафиксировать их), можно сравнивать по очереди *абсолютные значения коэффициентов корреляции* между Y и каждой из X-переменных.

Корреляция служит мерой силы такой взаимосвязи (о чем мы уже говорили в главе 11), однако почему следует использовать абсолютные значения? Вспомните, что корреляция, близкая к 1 или −1, указывает на сильную взаимосвязь, а корреляция, близкая к нулю, свидетельствует об отсутствии взаимосвязи. Абсолютное значение корреляции указывает на *силу* взаимосвязи, не определяя ее *направления*.

Множественная регрессия *делает поправку* на другие переменные, а коэффициент корреляции — нет.[11] Если вам требуется учитывать влияние других переменных, тогда следует пользоваться множественной регрессией. Если вам не нужно учитывать такую поправку, можно воспользоваться анализом корреляций.

Ниже приведены коэффициенты корреляции Y с каждой из X-переменных для примера с рекламными объявлениями в журналах. Например, корреляция между тарифом на размещение рекламы в журналах и медианой дохода читателей равна −0,167.

Корреляция со стоимостью страницы рекламы		
Читательская аудитория	Процент читателей-мужчин	Медиана дохода
0,872	−0,081	−0,167

С точки зрения взаимосвязи с тарифом на размещение рекламы в журналах (без поправки на другие X-переменные), размер читательской аудитории имеет наибольшее абсолютное значение корреляции, 0,872. Следующим по абсолютной величине корреляции является медиана дохода — $|-0{,}167| = 0{,}167$. Проценту читателей-мужчин соответствует наименьшее абсолютное значение корреляции — $|-0{,}081| = 0{,}081$. Все выглядит так, будто именно величина читательской аудитории практически полностью определяет величину тарифа на размещение рекламы в журналах. Действительно, ни одна из двух других переменных (сама по себе, без фиксации оставшихся переменных) не определяет значимую долю тарифа на размещение рекламы в журналах.

Множественная регрессия дает несколько иную картину, поскольку она позволяет контролировать значения других переменных. После внесения поправки на величину читательской аудитории коэффициент множественной регрессии для медианы дохода свидетельствует о значимом влиянии соответствующей переменной на величину рекламного тарифа. Это можно интерпретировать следующим образом. Поправка на величину читательской аудитории учитывает тот факт, что более высокие доходы сопутствуют меньшим читательским аудиториям. Влияние величины читательской аудитории нивелируется — остается лишь влияние дохода в чистом виде (которое проявляется благодаря тому, что снимается маскирующий эффект величины читательской аудитории).

Хотя коэффициенты корреляции указывают на *индивидуальные* взаимосвязи с Y, стандартизованные коэффициенты регрессии из множественной регрессии могут предоставить вам важную дополнительную информацию.

[11] Существует более совершенная статистическая концепция *коэффициента частной корреляции*, которую мы не будем описывать в этой книге. Такой коэффициент определяет корреляцию между двумя переменными с учетом поправки на одну или несколько дополнительных переменных.

12.2. Сложности и проблемы, связанные с множественной регрессией

К сожалению, на практике множественная регрессия не всегда позволяет получить результаты, о которых пишут в учебниках. В этом разделе приведен перечень потенциальных проблем и некоторые соображения по поводу того, как с ними справиться (в тех случаях, когда это возможно).

Существуют три основные разновидности проблем. Ниже приведен краткий обзор каждой из этих разновидностей, а затем следует более подробное их описание.

1. Проблема **мультиколлинеарности** возникает в тех случаях, когда некоторые из ваших объясняющих переменных (X) оказываются слишком схожими. Несмотря на то что эти переменные могут хорошо пояснять и прогнозировать Y (на что указывают высокое значение R^2 и значимый F-тест), отдельные коэффициенты регрессии плохо поддаются оценке. Это связано с тем, что мы не располагаем достаточной информацией, чтобы решить, какая (или какие) из переменных обеспечивает это объяснение. Одно из возможных решений состоит в том, чтобы удалить из уравнения некоторые из переменных с целью избавиться от сомнений. Другое решение заключается в том, чтобы переопределить какие-то из переменных (возможно, путем деления), чтобы отличать одну переменную от другой.
2. Проблема **выбора переменных** возникает в тех случаях, когда приходится иметь дело с пространным перечнем потенциально полезных объясняющих (независимых) X-переменных и необходимо решить, какие из этих переменных следует включать в уравнение регрессии. С одной стороны, если у вас слишком много X-переменных, лишние из них будут снижать качество результатов (возможно, по причине все той же мультиколлинеарности). Часть информации, содержащейся в данных, понапрасну расходуется на оценивание ненужных параметров. С другой стороны, если отбросить нужную X-переменную, снизится качество прогнозов, поскольку вы проигнорируете полезную информацию. Одно из возможных решений состоит в том, чтобы хорошенько подумать, *почему* важна та или иная X-переменная, чтобы быть уверенным в том, что каждая включаемая в рассмотрение переменная действительно выполняет важную функцию. Другой подход заключается в том, чтобы воспользоваться автоматической процедурой, которая старается отобрать наиболее важные переменные.
3. Проблема **неправильного выбора модели** связана с множеством различных потенциальных несоответствий между вашей конкретной задачей и моделью множественной линейной регрессии, которая является фундаментом и каркасом множественного линейного регрессионного анализа. Может получиться так, что ваша конкретная задача не соответствует условиям и допущениям модели линейной множественной регрессии. Анализируя данные, вы можете выявить некоторые потенциальные проблемы, связанные с нелинейностью, неравной изменчивостью и наличием резко отклоняющихся значений. Однако даже наличие подобных проблем еще ни о чем не говорит. Несмотря на то что гистограммы некоторых переменных могут

быть сильно скошенными (несимметричными), а некоторые диаграммы рассеяния могут быть нелинейными, модель множественной линейной регрессии и в таких случаях вполне может быть применима. Существует так называемая *диагностическая диаграмма*, которая помогает понять, действительно ли обнаруженная проблема является настолько серьезной, что ее необходимо как-то решать. Один из возможных вариантов решений заключается в создании новых X-переменных, которые формируются на основе существующих переменных, и/или преобразовании некоторых или всех этих переменных. Еще одна серьезная проблема возникает в случае, когда приходится иметь дело с *временным рядом*, применительно к которому допущение модели линейной множественной регрессии о независимости отдельных наблюдений не соблюдается. Проблема временных рядов не имеет простого решения, однако множественную регрессию можно выполнить, используя вместо исходных данных *процентные изменения* между различными временными периодами.

Мультиколлинеарность: не слишком ли схожи между собой объясняющие переменные?

Когда какие-то из объясняющих X-переменных слишком схожи между собой, у вас может возникнуть проблема *мультиколлинеарности*, поскольку множественная регрессия не в состоянии отличить влияние одной переменной от влияния другой переменной. Последствия мультиколлинеарности могут быть *статистическими* или *вычислительными*.

1. *Статистические* последствия мультиколлинеарности связаны с трудностями проведения статистических тестов для отдельных коэффициентов регрессии вследствие увеличения стандартных ошибок. Результатом может быть невозможность объявить ту или иную X-переменную значимой даже в том случае, если эта переменная (сама по себе) имеет сильную взаимосвязь с Y.
2. *Вычислительные* последствия мультиколлинеарности связаны с трудностями в организации вычислений на компьютере, вызванными "неустойчивостью вычислений". В крайних случаях компьютер может пытаться выполнить деление на нуль и, таким образом, неудачно завершить анализ данных. Хуже того, компьютер может завершить анализ и выдать бессмысленные и неверные результаты.[12]

Мультиколлинеарность может порождать проблемы, а может и не порождать их — все зависит от конкретных целей выполняемого вами анализа и степени мультиколлинеарности. Небольшая или средняя мультиколлинеарность обычно не представляет проблемы. Очень сильная мультиколлинеарность (например,

[12] Деление на нуль невозможно с математической точки зрения: например, результат выполнения 5/0 является неопределенным. Однако из-за небольших ошибок округления в процессе вычислений компьютер может разделить не 5 на 0, а 5,0000000000968 на 0,0000000000327. В этом случае, вместо того чтобы остановиться и сообщить об ошибке, компьютер использует в дальнейших вычислениях бессмысленный и огромный результат такого деления: 152 905 198 779,72.

включение одной и той же переменной дважды) всегда будет представлять проблему и может приводить к серьезным ошибкам (вычислительные последствия). К счастью, если вашей целью является в основном предсказание или прогнозирование Y, сильная мультиколлинеарность может не представлять серьезного препятствия, поскольку качественная программа множественной регрессии может и в этом случае делать оптимальные прогнозы Y (по методу наименьших квадратов), основанные на всех X-переменных. Однако если вы хотите использовать индивидуальные коэффициенты регрессии для выяснения того, как каждая из X-переменных влияет на Y, то статистические последствия мультиколлинеарности, по-видимому, вызовут определенные проблемы, ввиду того что эти влияния невозможно отделить друг от друга. В табл. 12.2.1 подытоживается влияние мультиколлинеарности на результаты регрессионного анализа.

Как выяснить, действительно ли существует проблема мультиколлинеарности? Один из простейших способов ответить на этот вопрос заключается в анализе обычных двумерных корреляций для каждой пары переменных.[13] *Корреляционная матрица* представляет собой таблицу, которая содержит коэффициенты корреляции для каждой пары переменных из вашей многомерной совокупности данных. Чем выше коэффициент корреляции между двумя X-переменными, тем больше мультиколлинеарность. Это объясняется тем, что высокая корреляция (близкая к 1 или −1) указывает на сильную связь и свидетельствует о том, что эти две X-переменные измеряют очень схожие характеристики, привнося тем самым в анализ "пересекающуюся" информацию.

Основной статистический результат мультиколлинеарности заключается в *росте стандартных ошибок некоторых или всех коэффициентов регрессии* (S_{b_i}). Это вполне естественно: если две X-переменные содержат "пересекающуюся" информацию, трудно определить влияние каждой из них в отдельности. Высокое значение стандартной ошибки приводит к тому, что компьютер сообщает вам приблизительно следующее: "Я вычислил для вас коэффициент регрессии, но результат неточный, поскольку трудно сказать, эта или какая другая переменная является определяющей". В результате доверительные интервалы для соответствующих коэффициентов регрессии значительно расширяются, а t-тесты вряд ли будут значимыми.

В случае сильной мультиколлинеарности может оказаться, что регрессия очень высоко значима (исходя из результатов F-теста), однако ни один из t-тестов для отдельных X-переменных значимым не является. Компьютер сообщает вам о том, что X-переменные, рассматриваемые как единая группа, весьма сильно влияют на Y, но практически невозможно определить важность какой-то конкретной переменной. Следует помнить, что t-тест для конкретной X-переменной измеряет ее влияние на Y при условии, что значения других переменных остаются неизменными. Таким образом, t-тест для переменной X_i выявляет только *дополнительную* информацию, привнесенную переменной X_i помимо

[13] К сожалению, исчерпывающий диагноз мультиколлинеарности оказывается гораздо сложнее, чем описываемый здесь способ, поскольку необходимо рассматривать все X-переменные одновременно, а не попарно. Полное техническое описание соответствующих методов можно найти, например, в книге Belsley D. A., Kuh E., and Welsch R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (New York: Wiley, 1980).

Таблица 12.2.1. Влияние мультиколлинеарности на регрессию

Степень мультиколлинеарности	Влияние на регрессионный анализ
Незначительная	Вообще не представляет проблемы
Средняя	*Как правило, не представляет проблемы*
Сильная	Статистические последствия: зачастую представляет собой проблему, если требуется оценить влияние отдельных *X*-переменных (т.е. коэффициенты регрессии); может не представлять проблемы, если цель заключается в предсказании или прогнозировании *Y*
Чрезвычайно сильная	Численные последствия: всегда представляет собой проблему; компьютерные вычисления могут даже оказаться неправильными из-за неустойчивости вычислений

той информации, которую несут другие X-переменные. Если какая-то другая переменная очень близка к X_i, тогда переменная X_i не привносит в регрессию значимо новую информацию.

Одно из решений заключается в том, чтобы проигнорировать те X-переменные, которые дублируют информацию, уже присутствующую в других X-переменных. Если, например, ваши X-переменные включают три *различные* измерения размера, попробуйте либо избавиться от двух из них, либо объединить все три переменные в единую меру размера (например, воспользовавшись их средним значением).

Другое решение заключается в том, чтобы переопределить некоторые из переменных с тем, чтобы каждая из X-переменных выполняла четкую, присущую только ей одной роль в определении Y. Распространенный способ применения этой идеи к группе близких друг к другу X-переменных заключается в том, чтобы взять для представления этой группы одну X-переменную (можно либо выбрать одну из этих X-переменных, либо сформировать из них индекс) и представить остальные переменные как относительные показатели (например, величина на единицу другого показателя), построенные из этой представляющей X-переменной. Например, можно представлять зависимость размера объема продаж (Y) с помощью численности населения (X_1) и общего дохода (X_2) для каждого региона. Однако эти переменные являются мультиколлинеарными (т.е. численность населения и общий доход — высоко коррелированные величины). Эту проблему можно решить, объясняя объем продаж (Y) с помощью численности населения (X_1) и размера дохода на одного человека (новая переменная X_2). В результате численность населения будет выполнять роль представляющей переменной, отражая общую величину территории, а доход, вместо того чтобы повторять уже известную нам информацию (о величине соответствующей территории), переопределяется и несет новую информацию (о благосостоянии людей).

Пример. *Прогнозирование рыночной стоимости на основании величины активов и количества служащих*

Какова рыночная стоимость фирмы и как она определяется? Рыночная стоимость фирмы представляет собой суммарную стоимость всех выпущенных в обращение акций, которую легко найти, умножив общее количество выпущенных в обращение акций на текущую цену акции. Она определяется соотношением

предложения и спроса на рынке ценных бумаг. Финансисты-теоретики говорят, что она представляет текущую стоимость (неопределенных, рисковых) будущих денежных потоков данной фирмы. Но как связана рыночная стоимость с другими характеристиками фирмы? Чтобы ответить на этот вопрос, можно воспользоваться множественной регрессией.

Рассмотрим информацию, представленную в табл. 12.2.2. В этой таблице сопоставляется рыночная стоимость фирмы (переменная Y — зависимая, "объясняемая") и две независимые ("объясняющие") X-переменные: стоимость активов, которыми *владеет фирма, и количество служащих фирмы*. Эти данные относятся к компаниям из перечня Fortune 500, которые производят компьютеры и офисное оборудование.

Нетрудно догадаться, что с этой совокупностью данных у нас наверняка возникнет проблема мультиколлинеарности, поскольку каждая X-переменная, в принципе, обозначает размер компании. Эти X-переменные несут схожую, "пересекающуюся" информацию, поскольку крупные компании являются крупными во всех отношениях: рыночной стоимости, величине активов и количестве служащих. И наоборот, небольшие компании, как правило, являются небольшими во всех отношениях. В табл. 12.2.3 подведены итоги результатов множественной регрессии.

Обратите внимание, что в соответствии с F-тестом регрессия является значимой. Свыше трех четвертей ($R^2 = 81{,}7\%$) вариации рыночной стоимости объясняются X-переменными как единой группой, и этот результат является высоко статистически значимым. Однако вследствие мультиколлинеарности ни одна отдельно взятая X-переменная значимой не является. Таким образом, X-переменные объясняют рыночную стоимость, но мы не можем сказать, какая именно из X-переменных оказывает наибольшее влияние.

Некоторую полезную информацию о мультиколлинеарности можно получить из корреляционной матрицы (табл. 12.2.4), которая отражает корреляции для каждой пары переменных нашей многомерной со-

Таблица 12.2.2. Компании из перечня Fortune 500, производящие компьютеры и офисное оборудование

	Рыночная стоимость (млн дол.), Y	Активы (млн дол.), X_1	Количество служащих, X_2
Intl. Business Machines	98 322	81 449	269 465
Hewlett-Packard	65 060	31 749	121 900
Compaq Computer	36 052	14 631	37 004
Xerox	31 829	27 732	91 400
Digital Equipment	7 101	9 693	54 900
Dell Computer	41 294	4 268	16 160
Sun Microsystems	16 614	4 697	21 500
Apple Computer	3 449	4 233	9 306
NCR	3 386	5 293	38 300
Gateway 2000	6 242	2 039	13 369
Pitney Bowes	14 036	7 893	29 901
Silicon Graphics	2 636	3 345	10 930
Data General	869	1 135	5 100
Intergraph	452	727	7 653

Данные получены из `http://www.pathfinder.com/fortune/fortune500/ind8.html`, осень 1998.



вокупности данных. Обратите внимание на чрезвычайно высокую корреляцию между двумя X-переменными: 0,991 между величиной активов и количеством служащих. Столь высокая корреляция свидетельствует о том, что по крайней мере с точки зрения чисел эти две X-переменные несут практически идентичную информацию. Нет ничего удивительного в том, что регрессионный анализ не делает различия между этими переменными.

Если бы мы сохранили только одну из двух X-переменных, то получили бы регрессию с очень высоко значимым t-тестом для этой переменной, независимо от того, какую из двух переменных X мы решили оставить. Иными словами, каждая из этих переменных сама по себе вносит весомый вклад в определение рыночной стоимости.

Если вы хотите сохранить всю информацию, содержащуюся в обеих X-переменных, одну из них можно использовать в качестве переменной, представляющей величину компании, а другую определить как некое отношение. Давайте выберем в качестве переменной, представляющей величину компании, ее активы, поскольку они указывают на фиксированные капиталовложения, необходимые соответствующей компании. После этого вторую переменную можно заменить на отношение количества служащих к величине активов (указывает количество служащих на миллион долларов активов). Теперь активы являются единственной *переменной, характеризующей величину компании, а другая переменная несет новую информацию об эф*фективности использования служащих. Новая совокупность данных представлена в табл. 12.2.5.

Таблица 12.2.3. Регрессионный анализ компаний, производящих компьютеры и офисное оборудование

Множественная регрессия для прогнозирования рыночной стоимости на основании активов и количества служащих.

Уравнение прогнозирования имеет следующий вид:

рыночная стоимость = 6998,40 + 1,68 (активы) – 0,1442 (количество служащих).

0,817	*R в квадрате*
13428	Стандартная ошибка оценки
14	Количество наблюдений
24,61	F-статистика
0,00009	p-значение

	Коэффициент	Нижняя граница 95% доверительного интервала	Верхняя граница 95% доверительного интервала	Стандартная ошибка	t	p
Константа	6998,40	–3660,98	17657,77	4843,00	1,45	0,176
Активы	1,68	–1,17	4,53	1,30	1,30	0,221
Количество служащих	–0,14	–1,01	0,72	0,39	–0,37	0,720

Таблица 12.2.4. Корреляционная матрица для компаний, производящих компьютеры и офисное оборудование

	Рыночная стоимость, Y	Активы, X_1	Количество служащих, X_2
Рыночная стоимость, Y	1,000	0,903	0,888
Активы, X_1	0,903	1,000	0,991
Количество служащих, X_2	0,888	0,991	1,000

Посмотрим теперь снова на корреляционную матрицу, представленную в табл. 12.2.6, и выясним, нет ли у нас проблем с мультиколлинеарностью. Эти корреляции выглядят намного лучше. Корреляция между X-переменными (–0,317) уже не является такой большой, как раньше, и она статистически незначима.

На что можно рассчитывать, получив результаты множественной регрессии? Регрессия по-прежнему должна быть значимой, а t-тест для активов на сей раз должен быть значимым по причине отсутствия "конкурирующих" переменных, характеризующих величину компании. Нам осталось разрешить следующую неопределенность: можно ли, располагая данными об активах, утверждать, что соотношение между количеством служащих и активами в значительной степени влияет на рыночную стоимость? Соответствующие результаты представлены в табл. 12.2.7.

Эти результаты подтверждают наши ожидания. Регрессия (F-тест) *является значимой, а t-тест* для активов теперь, когда нам удалось избавиться от сильной мультиколлинеарности, также является значимым. Кроме того, нам удалось установить, что другая переменная (количество служащих на миллион долларов активов) *значимой не является.*

Таблица 12.2.5. Определение новых X-переменных для компаний, производящих компьютеры и офисное оборудование; использование отношения количества служащих к размерам активов

	Рыночная стоимость, (млн дол.), Y	Активы, (млн дол.), X_1	Отношение количества служащих к размерам активов, X_2
Intl. Business Machines	98 322	81 449	3,308
Hewlett-Packard	65 060	31 749	3,839
Compaq Computer	36 052	14 631	2,529
Xerox	31 829	27 732	3,296
Digital Equipment	7 101	9 693	5,664
Dell Computer	41 294	4 268	3,786
Sun Microsystems	16 614	4 697	4,577
Apple Computer	3 449	4 233	2,198
NCR	3 386	5 293	7,236
Gateway 2000	6 242	2 039	6,557
Pitney Bowes	14 036	7 893	3,788
Silicon Graphics	2 636	3 345	3,268
Data General	869	1 135	4,493
Intergraph	452	727	10,527

Таблица 12.2.6. Корреляционная матрица для компаний, производящих компьютеры и офисное оборудование (используются новые X-переменные)

	Рыночная стоимость (млн дол.), Y	Активы (млн дол.),X_1	Отношение количества служащих к размерам активов, X_2
Рыночная стоимость, Y	1,000	0,903	–0,400
Активы, X_1	0,903	1,000	–0,317
Отношение количества служащих к размерам активов, X_2	–0,400	–0,317	1,000

Таблица 12.2.7. Регрессионный анализ компаний, производящих компьютеры и офисное оборудование (используются новые *X*-переменные)

Множественная регрессия для прогнозирования рыночной стоимости на основании объема активов и количества служащих на миллион долларов активов.

Уравнение прогнозирования имеет вид

рыночная стоимость = 14673,67 + 1,154 (активы) – 1655,524 (количество служащих на миллион долларов активов).

0,830	R в квадрате
12967	Стандартная ошибка оценки
14	Количество наблюдений
26,787	F-статистика
0,00006	p-значение

	Коэффициент	Нижняя граница 95% доверительного интервала	Верхняя граница 95% доверительного интервала	Стандартная ошибка	t	p
Константа	14673,67	–6637,27	35984,61	9682,45	1,52	0,158
Активы	1,15	0,77	1,54	0,18	6,57	0,000
Количество служащих на миллион долларов активов	–1655,52	–5413,44	2102,39	1707,38	–0,97	0,353

Очевидно, для этой небольшой группы (n = 14) крупных компаний, производящих компьютеры и офисное оборудование, большая доля вариации рыночной стоимости может объясняться объемом активов этих компаний. Более того, информация о людских ресурсах (количестве служащих) практически не содержит новой информации о рыночной стоимости этих процветающих компаний. Возможно, анализ более крупной выборки компаний позволил бы выявить влияние и этой переменной.

Выбор переменной: может быть, мы пользуемся "не теми" переменными?

Результаты статистического анализа в значительной мере зависят от имеющейся информации, т.е. от использованных для анализа данных. В частности, особое внимание следует обратить на выбор независимых ("объясняющих") X-переменных для множественного регрессионного анализа. Включение как можно большего числа X-переменных "просто так, на всякий случай" или потому, что "создается впечатление, будто каждая из них как-то влияет на Y" — далеко *не* лучшее решение. Поступая таким образом, вы обрекаете себя на возможные трудности при определении значимости для регрессии (F-тест), или — вследствие мультиколлинеарности, вызванной наличием избыточных переменных, — у вас могут возникнуть трудности при решении вопроса о значимости для некоторых отдельных коэффициентов регрессии.

Что происходит, когда вы включаете одну лишнюю, неуместную X-переменную? Значение R^2 в этом случае окажется несколько большим, так как несколько большую долю Y можно объяснить за счет случайности этой новой

переменной.[14] Однако F-тест значимости регрессии учитывает это увеличение, поэтому такое увеличение R^2 нельзя считать преимуществом.

На самом деле включение дополнительной X-переменной может принести небольшой или даже умеренный *вред*. Оценка того или иного неуместного параметра (в данном случае неуместного коэффициента регрессии) оставляет меньше информации для стандартной ошибки оценки, S_e. По техническим причинам следствием этого является менее мощный F-тест, который может не обнаружить значимость даже в том случае, когда X-переменные в генеральной совокупности на самом деле объясняют Y.

А что произойдет в случае, когда вы проигнорируете необходимую X- переменную? В результате из совокупности данных выпадет важная и полезная информация и ваше прогнозирование Y будет менее точным, чем в случае использования этой X-переменной. Стандартная ошибка оценки, S_e, в этом случае, как правило, оказывается больше (что указывает на большие ошибки прогнозирования), а R^2, как правило, оказывается меньшим (что указывает на объяснение меньшей доли вариации Y). Естественно, если вы проигнорируете критически важную X-переменную, то, возможно, F-тест для этой регрессии просто будет незначим.

Ваша задача в данном случае — включить ровно столько X-переменных, сколько нужно (т.е. не слишком много и не слишком мало), причем включить именно те X-переменные, которые необходимы. Если у вас есть сомнения, можно включить некоторые из X-переменных, относительно которых вы не уверены. В таком случае полезен субъективный метод (основанный на приоритетном перечне X-переменных). Существует также множество различных автоматических методов.

Классификация перечня *X*-переменных по приоритетам

Хороший способ определить круг важных X-переменных заключается в том, чтобы внимательно проанализировать решаемую задачу, имеющиеся данные и цели, которых вы хотите добиться. Затем необходимо составить список X-переменных, классифицированных по приоритетам. Сделать это можно следующим образом.

1. Выберите переменную Y, которую вам необходимое объяснить, понять или прогнозировать.
2. Выберите X-переменную, которая, как вам кажется, является наиболее важной в определении или объяснении Y. Если это вызывает у вас затруднения, поскольку *все* X-переменные кажутся вам одинаково важными, примите *волевое* решение.
3. Выберите самую важную среди оставшихся X-переменных, задав себе вопрос: "Принимая во внимание первую переменную, какая из оставшихся X-переменных несет больше *новой* информации, объясняющей поведение переменной Y?"

[14] Несмотря на то что R^2 в любом случае будет либо таким же, либо большим, существует аналогичная величина R^2, называемая *скорректированным* R^2, которая при включении ненужной X-переменной может оказаться либо большей, либо меньшей. Скорректированное R^2 увеличится лишь в том случае, если данная X-переменная объясняет больше, чем можно было бы ожидать вследствие всего лишь случайности от неуместной X-переменной. Скорректированное R^2 можно вычислить на основании обычного, нескорректированного значения R^2 по формуле $1 - (n - 1)(1 - R^2)/(n - k - 1)$.

4. Продолжайте выбирать по этому принципу самые важные из оставшихся X-переменных до тех пор, пока не классифицируете по приоритетам весь перечень X-переменных. На каждой стадии задавайте себе вопрос: "Принимая во внимание уже отобранные X-переменные, какая из оставшихся X-переменных несет больше *новой* информации, объясняющей поведение переменной Y?"

Затем вычислите регрессию, используя лишь те X-переменные из составленного вами списка, которые кажутся вам важнейшими. Вычислите еще несколько регрессий, включая в свой анализ некоторые из оставшихся X-переменных (или все эти переменные), и выясните, действительно ли они влияют на прогнозирование переменной Y. Наконец, выберите тот результат регрессии, который кажется вам наиболее полезным.

Несмотря на то что описанная процедура выглядит достаточно субъективной (поскольку зависит в основном от вашего субъективного мнения), ей присущи два важных преимущества. Во-первых, когда необходимо сделать выбор между двумя X-переменными, которые практически одинаково объясняют поведение переменной Y, окончательный выбор остается за вами (автоматизированная процедура может в этом случае сделать менее содержательный выбор). Во-вторых, тщательно классифицировав по приоритетам свои независимые X-переменные, вы можете глубже разобраться в исследуемой ситуации. Такое прояснение решаемой задачи может оказаться не менее полезным, чем результаты множественной регрессии!

Автоматизация процесса выбора переменных

Если вы не хотите тратить время на глубокие размышления над исследуемой ситуацией и предпочитаете автоматизировать процесс выбора X-переменных на основе имеющихся у вас данных, в вашем распоряжении есть немало способов достижения требуемого результата. К сожалению, "наилучшего" во всех отношениях способа автоматизации выбора переменных не существует. Ученые продолжают поиск такого способа, однако уже сейчас имеются достаточно хорошие автоматические методы, позволяющие получить относительно компактный перечень X-переменных, обеспечивающих вполне качественное прогнозирование Y.

Наилучшим методом автоматического выбора переменных является анализ *всех подмножеств* X-переменных. Если, например, вы располагаете тремя независимыми X-переменными, из которых вам нужно сделать свой выбор, тогда, как показано в табл. 12.2.8, необходимо исследовать восемь подмножеств этих переменных. Если вы располагаете десятью X-переменными, придется исследовать уже 1024 различных подмножеств.[15] Даже если у вас есть возможность вычислить такое количество регрессий, как вы узнаете, какое из подмножеств является наилучшим? Ученые-статистики предложили ряд технических методов, основанных на формулах, которые учитывают как дополнительную информацию, содержащуюся в более крупных подмножествах, так и дополнительные сложности оценки.[16]

[15] Общая формула имеет следующий вид: из k X-переменных можно сформировать 2^k подмножеств.

Один из широко практикуемых подходов называется *пошаговым выбором*. На каждом шаге переменная либо добавляется в список, либо удаляется из списка — в зависимости от своей "полезности". Этот процесс продолжается до тех пор, пока список переменных не стабилизируется. Эта процедура выполняется быстрее, чем анализ всех подмножеств переменных, но в некоторых случаях он может не привести к нужному результату. Вот некоторые подробности, касающиеся процедуры пошагового выбора.

1. *Инициализация*. Существует ли такая X-переменная, которая помогает объяснить Y? Если нет, остановить процедуру пошагового выбора и сообщить о том, что полезных X-переменных обнаружить не удается. Если же такую переменную удалось обнаружить, поместите эту наиболее полезную X-переменную в список (это одна из тех переменных, которые характеризуются наибольшим абсолютным значением корреляции с Y).
2. *Шаг включения переменной*. Проанализируйте все X-переменные, *не* включенные в список. Рассмотрите, в частности, ту X-переменную, которая в наибольшей мере *дополнительно* объясняет Y. Если это объяснение кажется вам достаточно важным, включите соответствующую X-переменную в список.
3. *Шаги удаления переменных*. Имеется ли в созданном списке такая X-переменная, которая в данный момент (после пополнения списка новыми переменными) кажется вам бесполезной? Если такая переменная в списке имеется, удалите ее, однако учтите, что, возможно, ее придется включить в список в дальнейшем. Продолжайте удалять бесполезные X-переменные до тех пор, пока их не останется в списке.
4. *Повторное выполнение до завершения процедуры*. Повторяйте действия, указанные в пп. 2 и 3 до тех пор, пока в список нечего будет добавить и нечего будет удалить.

Таблица 12.2.8. Список всех возможных подмножеств X-переменных для $k = 3$

1	Пустое множество (для прогнозирования Y можно использовать только $\bar{Y}$)
2	X_1
3	X_2
4	X_3
5	X_1, X_2
6	X_1, X_3
7	X_2, X_3
8	X_1, X_2, X_3

[16] Хорошей мерой для выбора наилучшего подмножества X-переменных в регрессии является C_p *статистика Мэллоуза* (Mallow's C_p statistic). Этот и другие подходы приведены в книге Draper N. R. and Smith H. *Applied Regression Analysis* (New York: Wiley, 1981), Chapter 6; и в книге Seber G. A. F. *Linear Regression Analysis* (New York: Wiley, 1977), Chapter 12.

Конечный результат процедуры пошагового выбора, как правило, представляет собой весьма полезный и достаточно компактный список независимых ("объясняющих") X-переменных, который можно использовать в множественном регрессионном анализе для объяснения Y.

Неправильный выбор модели: возможно, уравнение регрессии имеет неправильную форму?

Даже если вам удалось получить хороший список X-переменных, который содержит необходимую для объяснения Y информацию, это вовсе не значит, что все проблемы уже решены. Вы можете столкнуться с *неправильным выбором модели*, т.е. с неудачным представлением конкретной исследуемой ситуации с помощью модели множественной линейной регрессии. Ниже перечислены некоторые случаи неправильного выбора регрессионной модели.

1. Ожидаемая реакция Y на X-переменные может оказаться *нелинейной*. Иными словами, уравнение регрессии $a + b_1X_1 + b_2X_2 + ... + b_kX_k$ может неадекватно описывать истинную взаимосвязь между Y и X-переменными.
2. Может наблюдаться *неравная изменчивость* Y. Тем самым нарушается предположение о том, что стандартное отклонение, σ, в модели множественной линейной регрессии является постоянным независимо от значений X-переменных.
3. В данных не исключено наличие одного или нескольких *резко отклоняющихся значений* или *кластеров*, что может серьезно исказить оценки регрессии.
4. Вы можете иметь дело с *временным рядом*. Тогда случайная компонента модели множественной линейной регрессии уже не будет независимой от различных периодов времени. Вообще говоря, анализ временных рядов достаточно сложен (см. главу 14). Однако у вас есть возможность и в этом случае работать с множественной регрессией, пользуясь вместо исходных переменных соответствующими *процентными изменениями* переменных (между различными периодами времени).

Некоторые из этих проблем можно выявить, проанализировав все диаграммы рассеяния, построенные для каждой возможной пары переменных (например, в случае $k = 3$ можно построить шесть диаграмм рассеяния: $[X_1,Y]$, $[X_2,Y]$, $[X_3,Y]$, $[X_1, X_2]$, $[X_1, X_3]$, $[X_2, X_3]$). Чтобы анализ ситуации получился полным, все эти диаграммы рассеяния необходимо хотя бы кратко исследовать, чтобы постараться выявить потенциальные проблемы и трудности. При этом следует помнить, что эти диаграммы рассеяния могут преувеличивать необходимость коррекции. Например, зависимость Y от X_1 может оказаться нелинейной, что само по себе может не представлять для вас проблемы.

К счастью, существует более прямой метод, который зачастую позволяет выявить наличие серьезных проблем. *Диагностическая диаграмма* представляет собой отдельную диаграмму рассеяния остаточных значений в зависимости от прогнозируемых значений; такая диаграмма может позволить обнаружить наиболее

серьезные проблемы, включая нелинейность, неравную изменчивость и наличие выбросов (резко отклоняющихся значений). Таким образом, в качестве базовой информации можно использовать все диаграммы рассеяния для основных переменных, а затем воспользоваться диагностической диаграммой как основой для принятия решения о необходимости внесения в анализ тех или иных изменений.

Анализ данных с целью выявления нелинейности или неравной изменчивости

Анализируя все возможные диаграммы рассеяния (каждая диаграмма соответствует определенной паре переменных), можно исследовать большую часть структуры взаимосвязей между этими переменными. Такой анализ зачастую может дать весьма полезные сведения об изучаемой ситуации. Однако *всю* структуру взаимосвязей исследовать таким способом все же невозможно. Например, вы наверняка упустите из виду совместное влияние двух переменных на некоторую третью переменную, поскольку в каждом отдельном случае рассматриваете только две переменные.[17] Тем не менее основные диаграммы рассеяния дают немало полезной исходной информации.

Вернемся к нашему предыдущему примеру с рекламными объявлениями в журналах, когда величину тарифа на размещение рекламы в журналах (Y) необходимо объяснить величиной читательской аудитории (X_1), процентом читателей-мужчин (X_2) и средним доходом (X_3). Рассмотрим диаграммы рассеяния значений каждой из этих четырех переменных в зависимости от другой переменной (рис. 12.2.1–12.2.6).

Пригодится нам и корреляционная матрица, поскольку она позволяет получить общее представление о силе и направленности связи в каждой из этих диаграмм рассеяния (табл. 12.2.9).

Как можно было бы подвести итог этого исследования диаграмм рассеяния и анализа корреляций? Самая сильная связь наблюдается между размером читательской аудитории и величиной тарифа на размещение рекламы в журналах (рис. 12.2.1); достаточно сильная связь наблюдается также между величиной средних доходов и процентом читателей-мужчин (рис. 12.2.6). Из диаграмм рассеяния мы также узнаем, что журналы с наибольшей читательской аудиторией и самыми большими тарифами на размещение рекламы, как правило, ориентированы на группу читателей со средними доходами, что приводит к проявлениям неравной изменчивости (рис. 12.2.3, 12.2.5).

Представляет ли это проблему? Диагностическая диаграмма поможет вам разобраться, какие проблемы (если таковые действительно существуют) требуют особого внимания, и покажет, работает ли выбранное вами решение проблемы.

[17] Некоторые компьютерные системы могут поворачивать диаграмму разброса точек в реальном времени, что позволяет визуально анализировать трехмерные диаграммы сразу для трех переменных! Различные методы исследования многомерных данных рассматриваются в книге Chambers J. M., Cleveland W. S., Kleiner B., and Tukey P. A. *Graphical Methods for Data Analysis* (Boston: Duxbury Press, 1983).

Использование диагностической диаграммы для выяснения наличия проблем

Диагностическая диаграмма для множественной регрессии представляет собой диаграмму рассеяния ошибок прогнозирования (остатков) в зависимости от прогнозируемых значений; она позволяет выяснить, можно ли повысить качество прогнозирования, избавившись от соответствующих проблем в исходных данных.[18] Значения остатков, $Y - [a + b_1X_1 + b_2X_2 + ... + b_kX_k]$, откладываются по вертикальной оси, а прогнозируемые значения, $a + b_1X_1 + b_2X_2 + ... + b_kX_k$, — по горизонтальной. Поскольку методы решения проблем достаточно сложны (удаление резко отклоняющихся значений, преобразования данных и т.п.), проблему можно определить лишь в том случае, если она ясна и ярко выражена.

Внимание!

Не предпринимайте действий, если диагностическая диаграмма не дает ясного и четкого представления о проблеме.

Диагностическая диаграмма "читается" в основном так же, как и любая другая двумерная диаграмма рассеяния (см. главу 11). В табл. 12.2.10 показано, как интерпретировать полученные результаты.

Почему все происходит именно так, а не иначе? Остаточные значения представляют собой *необъясненные* ошибки прогнозирования Y, которые невозможно учесть с помощью модели множественной линейной регрессии, включающей X-переменные. Прогнозируемые значения представляют собой *текущее объяснение* исходя из X-переменных. Если в диагностической диаграмме наблюдается определенная достаточно сильная взаимосвязь, текущее объяснение можно и нужно улучшить, внеся изменения, учитывающие эту видимую взаимосвязь.

На рис. 12.2.7 показана диагностическая диаграмма, относящаяся к примеру с рекламными объявлениями в журналах. Здесь величина тарифа на размещение рекламы в журналах (Y) объясняется величиной читательской аудитории (X_1), процентом читателей-мужчин (X_2) и средним доходом (X_3). На диаграмме виден наклон; в нижнем правом углу отчетливо выделяются три резко отклоняющихся значения (выброса). Эти резко отклоняющиеся значения могут существенно ухудшать качество прогнозирования для остальных данных; если нам удастся каким-то образом избавиться от них, мы, возможно, повысим качество уравнения прогнозирования.

Гистограмма величины читательской аудитории, показанная на рис. 12.2.8, демонстрирует очень большую асимметрию, тогда как в гистограммах других переменных (эти гистограммы не показаны) такая асимметрия отсутствует. Несмотря на то что преобразовывать X-переменные лишь по причине асимметрии нет большой необходимости, мы все же посмотрим, что произойдет, если преобразовать переменную величины читательской аудитории (X_1).

[18] *Прогнозируемые значения* иногда называют *подогнанными значениями* (или *вычисленными значениями*).

На рис. 12.2.9 показана гистограмма для натуральных логарифмов величины читательской аудитории, $\log X_1$ (можно воспользоваться функцией LN в Excel).[19] В результате такого преобразования нам в основном удалось избавиться от асимметрии распределения. Теперь посмотрим, улучшает ли такое преобразование величины читательской аудитории результат регрессии.

Таблица 12.2.9. Корреляционная матрица для данных о размещении рекламных объявлений в журналах

	Рекламный тариф, Y	Читательская аудитория, X_1	Процент читателей-мужчин, X_2	Медиана дохода, X_3
Рекламный тариф, Y	1,000	0,872	–0,081	–0,167
Читательская аудитория, X_1	0,872	1,000	–0,134	–0,353
Процент читателей-мужчин, X_2	–0,081	–0,134	1,000	0,564
Медиана дохода, X_3	–0,167	–0,353	0,564	1,000

Таблица 12.2.10. Как интерпретировать диагностическую диаграмму зависимости значений остатков от прогнозируемых значений для множественной регрессии

Структура в диагностической диаграмме	Интерпретация
Взаимосвязь отсутствует; совершенно случайное распределение, без наклона	Вам повезло: никаких проблем не обнаружено. Возможно, некоторые улучшения и необходимы, но диагностическая диаграмма не может их определить
Линейная взаимосвязь с наклоном	Невозможна сама по себе, поскольку найденное методом наименьших квадратов уравнение регрессии, скорее всего, уже учитывает любую чисто линейную взаимосвязь
Линейная взаимосвязь с наклоном и резко отклоняющимся значением (значениями)	Резко отклоняющееся значение (значения) исказило коэффициенты регрессии и прогнозы. Прогнозы для той части данных, которые "ведут себя хорошо", можно улучшить, если вы чувствуете, что резко отклоняющиеся значения можно контролировать (возможно, с помощью некоторого преобразования) или проигнорировать*
Нелинейная взаимосвязь, как правило, U-образной формы или повернутой U-образной формы	В данных обнаружена нелинейная взаимосвязь. Качество ваших прогнозов можно повысить, либо выполнив преобразование, либо включив дополнительную переменную, либо воспользовавшись нелинейной регрессией
Неравная изменчивость	Оценка уравнения прогнозирования является недостаточно эффективной. Слишком большое значение имеет менее надежная часть данных, а наиболее надежная часть данных не имеет должного значения. Эту проблему можно решить, преобразовав Y (возможно, наряду с некоторыми из X-переменных)

*Преобразование никогда не следует применять *исключительно* для контроля резко отклоняющихся значений. Вы можете выполнить преобразование с целью снижения чрезмерной асимметрии и обнаружить, что прежние резко отклоняющиеся значения уже не являются таковыми.

[19] Например, величина читательской аудитории журнала *Audubon* равна 1 645 (в тысячах человек). Натуральный логарифм (иногда обозначаемый как ln) числа 1 645 равен 7,405.

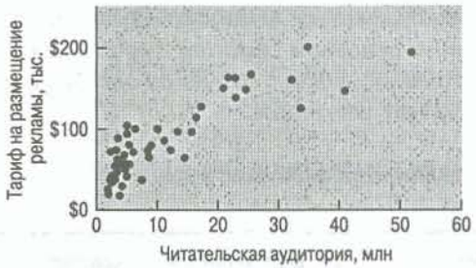


Рис. 12.2.1. Диаграмма рассеяния Y (тариф на размещение рекламы в журналах) в зависимости от X_1 (величина читательской аудитории) демонстрирует сильную взаимосвязь увеличивающего типа

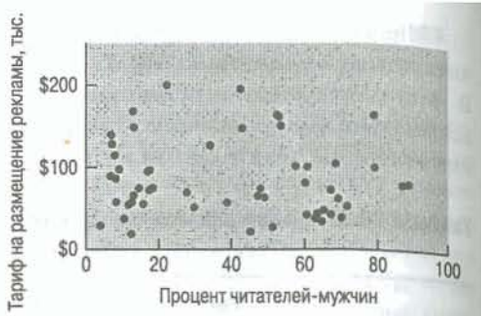


Рис. 12.2.2. Диаграмма рассеяния Y (тариф на размещение рекламы в журналах) в зависимости от X_2 (процент читателей-мужчин) демонстрирует практически полное отсутствие структуры

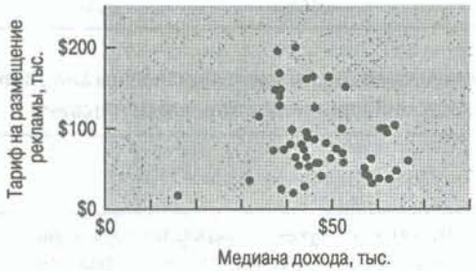


Рис. 12.2.3. Диаграмма рассеяния Y (тариф на размещение рекламы в журналах) в зависимости от X_3 (медиана дохода), на первый взгляд, демонстрирует практически полное отсутствие структуры. Более пристальное ее исследование, однако, указывает на существование некоторой тенденции к использованию низкого тарифа на размещение рекламы в журналах в двух крайних точках (низкие и высокие доходы), а также к высокой изменчивости рекламного тарифа для группы читателей со средними доходами. Может оказаться достаточно затруднительным использовать высокие рекламные тарифы на нижнем конце шкалы доходов (поскольку такие читатели просто мало тратят), а также на ее верхнем конце (поскольку читателей с такими доходами слишком мало)

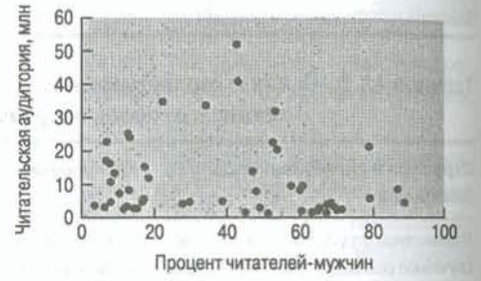


Рис. 12.2.4. Диаграмма рассеяния X_1 (величина читательской аудитории) в зависимости от X_2 (процент читателей-мужчин) демонстрирует практически полное отсутствие структуры (в лучшем случае такая структура едва просматривается)

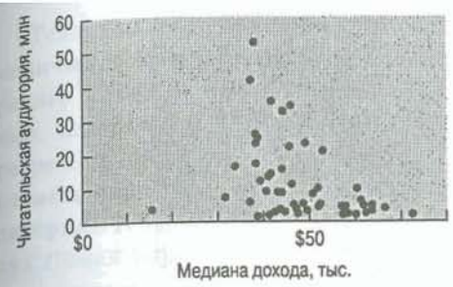


Рис. 12.2.5. Диаграмма рассеяния X_1 (величина читательской аудитории) в зависимости от X_3 (медиана дохода) показывает, что журналы, располагающие большой читательской аудиторией, как правило, ориентируются на читателей со средними доходами, но внутри этой группы наблюдается значительная изменчивость. Крайним значениям (высокие и низкие доходы) обычно соответствует незначительная по величине аудитория

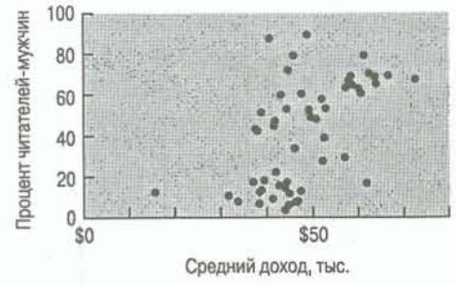


Рис. 12.2.6. Диаграмма рассеяния X_2 (процент читателей-мужчин) в зависимости от X_3 (медиана дохода) свидетельствует о существовании связанных с полом различий в уровне дохода. Журналы, ориентированные на читателей с высоким уровнем доходов, как правило, располагают более высоким процентом читателей-мужчин; среди читателей журналов, ориентированных на читателей с низким уровнем доходов, как правило, встречается больше женщин. У журналов, ориентированных на читателей со средним уровнем доходов, наблюдается большой разброс по признаку пола читателей

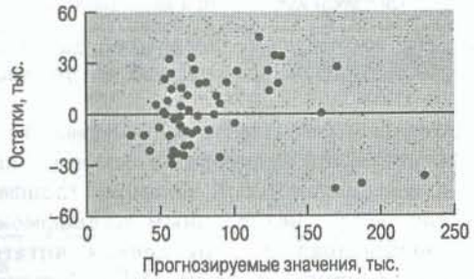


Рис. 12.2.7. Эта диагностическая диаграмма демонстрирует некоторую возможно необъясненную структуру в остатках: обратите внимание на наклон вверх основной части диаграммы рассеяния, действительной причиной которого может быть наличие трех резко отклоняющихся значений внизу справа. Это — диагностическая диаграмма множественной регрессии основных переменных для объяснения тарифа на размещение рекламы в журналах (Y) с помощью величины читательской аудитории (X_1), процента читателей-мужчин (X_2) и медианы дохода (X_3)

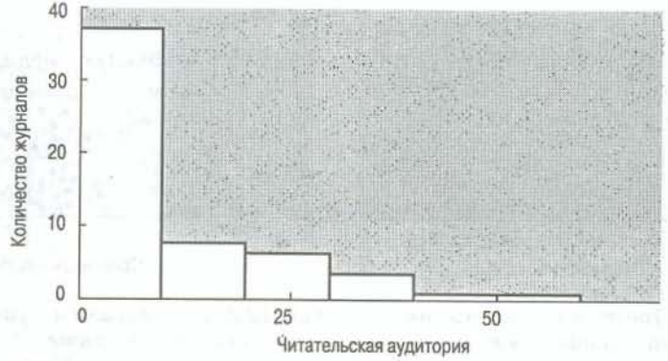


Рис. 12.2.8. Гистограмма размера читательской аудитории (X_1) демонстрирует очень большую асимметрию

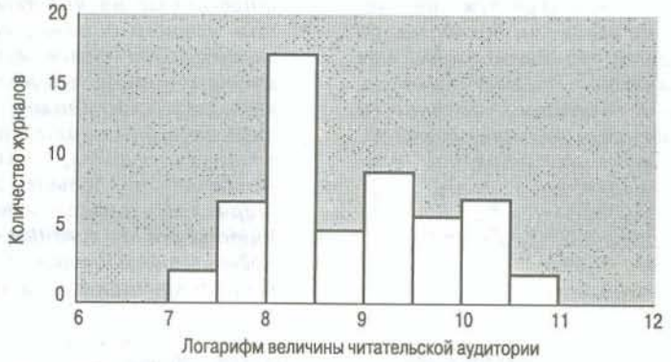


Рис. 12.2.9. На гистограмме логарифма размера читательской аудитории асимметрия отсутствует

В табл. 12.2.11 представлены результаты множественной регрессии после преобразования величины читательской аудитории с помощью натурального логарифма. Теперь переменные представляют собой величину тарифа на размещение рекламы в журналах (Y), объясняемую натуральным логарифмом величины читательской аудитории (новая переменная X_1), процентом читателей-мужчин (X_2) и медианой дохода (X_3). Можно отметить несколько небольших улучшений: улучшилось (т.е. увеличилось, что свидетельствует о лучшем объяснении Y) с 78,7 до 80,5% значение R^2, а стандартная ошибка оценки несколько уменьшилась с $21 578 до $20 662. Можно сказать, что выполненное нами преобразование позволяет лучше понять и прогнозировать тариф на размещение рекламы в журналах.

Диагностическая диаграмма для этой регрессии, представленная на рис. 12.2.10, определенно отличается от диагностической диаграммы для исходных данных (рис. 12.2.7): в частности, три потенциально резко отклоняющихся значения теперь находятся среди остальных данных. Однако возникла новая проблема: в данных просматривается определенная нелинейность (с обеих сторон диаграммы наблюдается некоторый подъем). Здесь есть определенный потенциал для улучшения соответствия между уравнением и данными.

Теперь давайте попытаемся преобразовать все переменные, которые измеряют количество (т.е. тариф на размещение рекламы в журналах, медиану дохода и размер читательской аудитории), одинаковым способом — с помощью натуральных логарифмов.[20] В табл. 12.2.12 представлены результаты множественной регрессии после преобразования с помощью натурального логарифма тарифа на размещение рекламы в журналах, медианы дохода и величины читательской аудитории. Теперь мы имеем логарифм тарифа на размещение рекламы в журналах (новая переменная Y), который объясняется с помощью логарифма величины читательской аудитории (новая переменная X_1), процента читателей-мужчин (X_2) и логарифма медианы дохода (новая переменная X_3). Значение R^2 повышается весьма несущественно, что свидетельствует о незначительном общем улучшении. Стандартное отклонение оценки теперь представлено в логарифмической шкале для тарифа на размещение рекламы в журналах, и поэтому его невозможно непосредственно сравнивать с предыдущими значениями.[21] Диагностическая диаграмма подскажет, насколько полезными оказались выполненные преобразования.

Диагностическая диаграмма для этой регрессии, показанная на рис. 12.2.11, свидетельствует о том, что с проблемой нелинейности нам удалось справиться, преобразовав с помощью логарифма величину тарифа на размещение рекламы в журналах, величину читательской аудитории и медиану дохода.

Таблица 12.2.11. Результаты множественной регрессии после логарифмирования размера читательской аудитории

Уравнение регрессии имеет следующий вид:

тариф на размещение рекламы = –370068 + 45730 **log** (аудитория) + 6 (процент мужчин) + 0,823 (доход).

Независимая переменная	Коэффициент	Стандартное отклонение	t	p
Константа	–370068	37101	–9,97	0,000
log Аудитория	46730	3281	14,24	0,000
Процент мужчин	6,2	131,5	0,05	0,963
Доход	0,8232	0,3516	2,34	0,023

S = 20662 R-квадрат = 80,5% R-квадрат(коррект.) = 79,3%

[20] Если какая-либо переменная (в другой ситуации) содержит как положительные, так и отрицательные значения, выполнение преобразования может вызвать некоторые затруднения и логарифм в этом случае невозможно использовать, поскольку для нулевого и отрицательных значений он не определен. В некоторых ситуациях можно попытаться переопределить переменную таким образом, чтобы она всегда была положительной. Если, например, она представляет собой прибыль (= доход – затраты), можно попытаться использовать вместо такой переменной отношение "доход/затраты". Тогда соответствующий логарифм примет следующий вид: log(доход/затраты) = log(доход) – log(затраты). Этот логарифм можно представить себе как отображение прибыли на процентной, а не на абсолютной "долларовой" шкале.

[21] Интерпретация результатов множественной регрессии в случае использования логарифмов будет изложена ниже в этой главе.

Окончание табл. 12.2.11

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	3	89752141196	29917380399	70,07	0,000
Ошибка остатка	51	21773821157	426937670		
Итого	54	1,11526E+11			

Источник	DF	Seq SS
log Аудитория	1	86384353889
Процент мужчин	1	1027578479
Доход	1	2340208827

Необычные наблюдения

Наблюдения	log Аудитория	Тариф на рекламу	Соответствие	Стандартное отклонение соответствия	Остаток	Стандартизованный остаток
2	10,5	198000	153262	5754	44738	2,25R
20	9,6	63750	111463	4158	–47713	–2,36R
48	8,2	17100	25386	11147	–8286	–0,48X
54	8,1	87500	45511	5541	41989	2,11R

R помечает наблюдение с большим стандартизованным остатком.

X помечает наблюдение, X-значение которого обеспечивает ему существенное влияние.

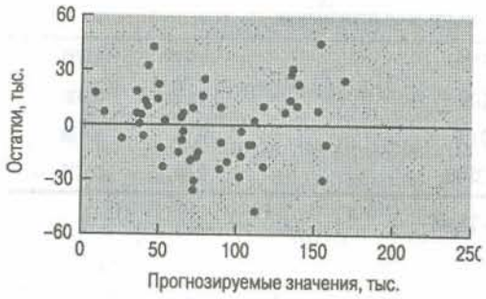


Рис. 12.2.10. Диагностическая диаграмма после логарифмирования величины читательской аудитории. В данном случае проблему может представлять нелинейность (наблюдается тенденция к подъему кривой на обоих концах)

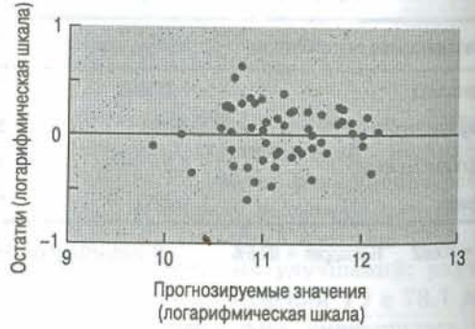


Рис. 12.2.11. На этом рисунке представлена прекрасная диагностическая диаграмма — без каких-либо потенциальных проблем в данных. После логарифмирования тарифа на размещение рекламы в журналах (Y), величины читательской аудитории (X_1) и медианы дохода (X_3) какой-либо взаимосвязи не наблюдается. Только процент читателей-мужчин (X_2) остался не преобразованным

Таблица 12.2.12. Результат множественной регрессии после логарифмирования тарифа на размещение рекламы в журналах, величины читательской аудитории и медианы дохода

Уравнение регрессии имеет следующий вид:

log Тариф на размещение рекламы = –3,44 + 0,578 (log Аудитория) – 0,00163 (процент мужчин) + 0,890 (log Доход).

Независимая переменная	Коэффициент	Стандартное отклонение	t	p
Константа	–3,441	2,011	–1,71	0,093
log Аудитория	0,57847	0,04023	14,38	0,000
Процент мужчин	–0,001635	0,001613	–1,01	0,316
log Доход	0,8897	0,1793	4,96	0,000

S = 0,2603 R-квадрат = 80,5% R-квадрат (коррект.) = 79,4%

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	3	14,2737	4,7579	70,22	0,000
Ошибка остатка	51	3,4557	0,0678		
Итого	54	17,7294			

Источник	DF	Seq SS
log Аудитория	1	12,4115
Процент мужчин	1	0,1945
log Доход	1	1,6677

Необычные наблюдения

Наблюдения	log Аудитория	log Тариф на рекламу	Соответствие	Стандартное отклонение соответствия	Остаток	Стандартизованный остаток
9	8,3	10,2421	10,8394	0,0705	–0,5973	–2,38R
48	8,2	9,7468	9,8707	0,1889	–0,1239	–0,69X
54	8,1	11,3794	10,7636	0,0707	0,6158	2,46R
55	8,0	11,2019	10,6940	0,0629	0,5079	2,01R

R помечает наблюдения с большим стандартизованным остатком.

X помечает наблюдение, X-значение которого обеспечивает ему существенное влияние.

Использование процентных изменений для моделирования экономических временны*х* рядов

Одно из предположений относительно модели множественной линейной регрессии заключается в том, что случайная компонента (ε) не зависит от конкретных значений данных. Когда вы имеете дело с данными временного ряда, это

предположение часто необоснованно, поскольку изменения при переходе от одного периода к следующему, как правило, весьма незначительны; тем не менее, за более длительные периоды времени возможны более значительные изменения.

Еще одним способом понять эту проблему является признание того факта, что многие экономические временные ряды с течением времени возрастают: например, валовой национальный продукт, доход после уплаты налогов и, смеем надеяться, объем продаж вашей фирмы. Множественная регрессия одной такой переменной (Y) на другие (X-переменные) будет характеризоваться высоким значением R^2, что предполагает наличие сильной связи. Но если каждый такой временной ряд возрастает с течением времени *самостоятельно*, присущим лишь ему одному способом и безотносительно остальных, это может привести к заблуждению. На самом деле вывод о наличии значимой связи можно сделать лишь в том случае, если *способ* увеличения Y с течением времени можно прогнозировать на основе увеличения X-переменных.

Один из способов решения этой проблемы заключается в том, чтобы работать с *процентными изменениями* каждой переменной, которые определяются соотношением (текущее – предыдущее)/предыдущее и представляют собой процент приращения соответствующей переменной за один период. Поступая так, вы ничего не теряете, поскольку проблему прогнозирования можно рассматривать либо как прогнозирование *изменения* по отношению к текущему уровню Y, либо как прогнозирование *будущего уровня* Y.

Представим себе систему, которая в каждый период времени пребывает в состоянии относительного равновесия, но претерпевает определенные изменения при переходе от одного периода к следующему. В действительности вас интересует, как воспользоваться информацией об X-переменных для прогнозирования очередного значения интересующей вас переменной Y. Одна из проблем заключается в том, что ваша совокупность данных представляет прошлую "историю" X-значений, которые уже не имеет смысла рассматривать как возможности. Работая с процентными изменениями, вы делаете эту прошлую "историю" более пригодной для своего текущего опыта. Иными словами, несмотря на то что объем продаж вашей фирмы наверняка существенно отличается от того, каким он был пять лет назад, процентные изменения объемов продаж между одним годом и следующим могут не очень существенно отличаться на протяжении длительного периода времени. Либо, если вы используете валовой национальный продукт (ВНП) для прогнозирования какой-то другой переменной, то, несмотря на то что абсолютное значение ВНП наверняка не будет таким же, каким оно было 10 лет назад, можно вполне рассчитывать примерно на такой же прирост (процентное изменение) ВНП.

Это можно представить себе следующим образом. Система в состоянии равновесия может обнаруживать тенденцию примерно к одинаковому изменению на протяжении длительного времени — несмотря на то, что с течением времени ее состояние может существенно измениться.

Может оказаться, что ваше значение R^2 пострадает, если вы воспользуетесь процентными изменениями вместо исходных значений данных. В некоторых случаях регрессия может потерять свою значимость. Поначалу это может "произвести плохое впечатление" (ну кто же не любит больших значений R^2!), но более внимательный анализ нередко показывает, что первоначальное значение R^2 было чересчур оптимистическим, а новое, меньшее значение оказывается ближе к истине.

Пример. Прогнозирование дивидендов

Как американские фирмы устанавливают свои дивиденды? На первый взгляд, можно прийти к выводу, что дивиденды в точности соответствуют ежегодному уровню продаж товаров недлительного пользования. Если, однако, пользоваться методом процентных изменений, то можно прийти к выводу, что объяснить изменения дивидендов не так-то просто.

Обратите внимание, что каждый из столбцов в табл. 12.2.13 свидетельствует об общем увеличении соответствующих переменных с течением времени. Следует, таким образом, предположить наличие сильной корреляции между этими переменными, поскольку высокие значения одной из них соответствуют высоким значениям других. Именно это мы и наблюдаем в корреляционной матрице, показанной в табл. 12.2.14.

Ничего удивительного нет и в чрезвычайно высоком значении R^2, указывающем на то, что впечатляющие 94,7% вариации дивидендов объясняются объемами продаж товаров недлительного и долговременного пользования. Однако это *не имеет ничего общего с действительностью*! Точнее говоря, в историческом контексте это, конечно, правильно, однако для прогнозирования будущих уровней дивидендов это мало что дает.

В табл. 12.2.15 показаны процентные изменения этих переменных. Например, величина изменения дивидендов в 1991 г. составляет (163 – 152)/152 = 7,24%. (Обратите внимание, что данные за 1990 г. отсутствуют, поскольку в исходной совокупности нет данных за предшествующий год.) Матрица корреляций, представленная в табл. 12.2.16, свидетельствует о значительно более умеренной связи между изменениями этих переменных при переходе от одного года к следующему. В сущности, при столь малом размере выборки (n = 6 для процентных изменений) ни одна из этих парных корреляций даже не является значимой. Величина R^2 для множественной регрессии процентных изменений снизилась до 24,7%,

Таблица 12.2.13. Дивиденды, объемы продаж товаров недлительного и долговременного пользования

Год	Дивиденды (млрд дол.), Y	Объемы продаж товаров недлительного пользования (млрд дол.), X_1	Объемы продаж товаров долговременного пользования (млрд дол.), X_2
1990	152	1 454	1 357
1991	163	1 457	1 304
1992	170	1 500	1 390
1993	197	1 524	1 490
1994	211	1 601	1 660
1995	227	1 715	1 804
1996	244	1 820	1 934

Данные взяты из таблиц 877 и 881 Бюро переписи населения США, *Statistical Abstract of the United States: 1997* (117th edition) Washington, DC, 1997.

Таблица 12.2.14. Матрица корреляций для дивидендов, объемов продаж товаров недлительного и долговременного пользования

	Дивиденды, Y	Товары недлительного пользования, X_1	Товары долговременного пользования, X_2
Дивиденды, Y	1,000	0,955	0,973
Товары недлительного пользования, X_1	0,955	1,000	0,986
Товары долговременного пользования, X_2	0,973	0,986	1,000

а F-тест уже не является значимым. Это указывает на то, что изменения в уровне объемов продаж товаров недлительного и долговременного пользования не позволяют пояснить изменение величины дивидендов при переходе от одного года к следующему.

С экономической точки зрения регрессионный анализ с помощью процентных изменений можно считать более оправданным. Колебания уровня дивидендов в экономике представляют собой сложный процесс, включающий взаимодействие множества факторов. Вследствие особенностей американской налоговой системы и явной неприязни инвесторов к внезапным изменениям в уровне дивидендов вряд ли можно рассчитывать на то, что колебания уровня дивидендов будут почти полностью объясняться исключительно объемами продаж.

Таблица 12.2.15. Годовые процентные изменения дивидендов, объемов продаж товаров недлительного и долговременного пользования

Год	Дивиденды (годовое изменение), Y, %	Объемы продаж товаров недлительного пользования (годовое изменение), X_1, %	Объемы продаж товаров долговременного пользования (годовое изменение), X_2, %
1990	–	–	–
1991	7,24	0,21	–3,91
1992	4,29	2,95	6,60
1993	15,88	1,60	7,19
1994	7,11	5,05	11,41
1995	7,58	7,12	8,67
1996	7,49	6,12	7,21

Таблица 12.2.16. Матрица корреляций для процентных изменений дивидендов, объемов продаж товаров недлительного и долговременного пользования

	Дивиденды, Y	Товары недлительного пользования, X_1	Товары долговременного пользования, X_2
Дивиденды, Y	1,000	–0,287	0,077
Товары недлительного пользования, X_1	–0,287	1,000	0,718
Товары долговременного пользования, X_2	0,077	0,718	1,000

12.3. Нелинейные взаимосвязи и неравная изменчивость

Методы множественной регрессии, которые мы до сих пор обсуждали, основываются на линейной модели множественной регрессии, которая характеризуется *постоянной изменчивостью*. Если вашей совокупности данных не присуща подобная линейная взаимосвязь, на что может указывать диагностическая диаграмма, которую мы исследовали выше, у вас есть три варианта действий. Первые два предусматривают применение множественной регрессии и описаны в настоящем разделе.

1. *Преобразовать некоторые (или все) переменные.* Преобразуя одну или несколько переменных (например, с помощью логарифмов), иногда удается получить новую совокупность данных, характеризующуюся линейной взаимосвязью. Помните, что логарифмы можно использовать для преобразования лишь положительных чисел. Если ваша совокупность данных характеризуется неравной изменчивостью, с этой проблемой можно справиться путем преобразования Y и (возможно) некоторых из X-переменных.
2. *Ввести новую переменную.* Ввод дополнительной, необходимой переменной X (например, X_1^2, "X_1 в квадрате") иногда позволяет получить линейную взаимосвязь между Y и новой совокупностью X-переменных. Такой метод может быть удачным, когда вам требуется найти *оптимальное* значение Y, например максимизировать прибыль или выпуск продукции. В других ситуациях можно использовать произведения переменных (например, определив $X_5 = X_1 \times X_2$), чтобы уравнение регрессии отражало *взаимодействие* между этими двумя переменными.
3. *Использовать нелинейную регрессию.* Иногда в данных может присутствовать важная нелинейная взаимосвязь (возможно, имеющая под собой определенное теоретическое обоснование), которую необходимо оценить непосредственно. В таких случаях можно воспользоваться более сложными методами *нелинейной регрессии* — если нам известны вид этой взаимосвязи и вид случайности.[22]

Преобразование взаимосвязи в линейную форму: интерпретация результатов

Выполняя преобразование своих данных, следует иметь в виду одну полезную рекомендацию. Чтобы избежать чрезмерного усложнения задачи, пытайтесь использовать одно и то же преобразование для всех переменных, которые измеряются в одних и тех же единицах. Если, например, вы логарифмируете объем продаж (который измерен в долларах или тысячах долларов), вам, вероятно, следует преобразовать таким же способом и все другие переменные, измеренные в долларах. При этом долларовые величины для всех соответствующих переменных будут измеряться по процентной шкале, а не по абсолютной "долларовой" шкале (именно в этом и заключается результат логарифмирования).

Правило соответствия для преобразования многомерных данных

Ко всем переменным, измеренным в одинаковых базовых единицах, желательно применять одно и то же преобразование.

Если вы выполняете множественный регрессионный анализ после преобразования всех или некоторых из переменных, то некоторые результаты могут требовать новой интерпретации. В этом разделе будет показано, как интерпретировать результаты множественного регрессионного анализа, когда либо (1) Y не подвергает-

[22] Введение в нелинейную регрессию можно найти в книге Draper N. R. and Smith H. *Applied Regression Analysis*, 2nd ed. (New York: Wiley, 1981), Chapter 10.

ся преобразованиям (т.е. преобразуются лишь некоторые или все X-переменные), либо (2) Y преобразуется с помощью натурального логарифма (независимо от того, преобразуются все или некоторые из X-переменных). Переменная Y играет особую роль, поскольку именно ее мы пытаемся прогнозировать. Поэтому преобразование Y переопределяет смысл ошибки прогнозирования.

Табл. 12.3.1 содержит интерпретацию основных результатов компьютерных вычислений: коэффициента детерминации, R^2; стандартной ошибки оценки, S_e; коэффициентов регрессии, b_i; и проверки значимости для b_i в случае использования преобразований.[23] Включена также процедура нахождения с помощью уравнения регрессии прогнозируемых значений Y.

Значение R^2 имеет одну и ту же базовую интерпретацию, независимо от того, как именно вы преобразуете свои переменные.[24] Это значение говорит о том, какая доля изменчивости вашего текущего Y (в любой — преобразованной или не преобразованной — форме) объясняется текущей формой X-переменных.

Стандартная ошибка оценки, S_e, имеет разную интерпретацию в зависимости от того, выполнялось ли преобразование Y. Если переменная Y не преобразовывалась, применяется обычная интерпретация (типичная величина ошибок прогнозирования), поскольку прогнозируется сама переменная Y. Однако если в регрессионном анализе используется log Y, то Y фигурирует в регрессии в процентах, а не в абсолютных значениях соответствующих единиц измерения. Подходящей мерой относительной изменчивости, в соответствии с материалом главы 5, является *коэффициент вариации*, поскольку та же изменчивость процентов будет как для высоких, так и для малых прогнозируемых значений Y. Формула для этого коэффициента вариации в табл. 12.3.1 базируется на теории логнормального распределения.[25]

Коэффициенты регрессии, b_i, если переменная Y не подвергалась преобразованиям, имеют обычную интерпретацию: они показывают ожидаемое влияние увеличения X_i на Y, причем единица увеличения X_i зависит от того, какому преобразованию подвергалась X_i. Если же переменная Y подвергалась преобразованиям, то b_i указывает на изменение в *преобразованной переменной* Y. Если вы использовали и логарифм переменной Y, и логарифм X_i, то b_i имеет специальную экономическую интерпретацию *эластичности*. **Эластичность** Y по отношению к X_i представляет собой ожидаемое процентное изменение Y, связанное с увеличением X_i на 1% при неизменных значениях других X-переменных; эластичность оценивается с помощью коэффициента регрессии из уравнения, где используются натуральные логарифмы как Y, так и X_i. Таким образом, эластичность — это почти то же самое, что и коэффициент регрессии, за исключением того, что изменения выражаются в процентах, а не в исходных единицах измерения.

Проверка значимости для коэффициента регрессии b_i сохраняет свою обычную интерпретацию для любых приемлемых вариантов преобразования. Главный вопрос заключается в следующем: оказывает ли X_i ощутимое влияние на Y (при условии, что другие X-переменные остаются неизменными) или Y ведет себя случайно по отношению к X_i? Поскольку ответом на этот вопрос является не подробное описание, а лишь "да" или "нет", основной предмет проверки остается тем же, независимо от того, выполняем мы логарифмическое преобразование или нет. Разумеется, в каждом отдельном случае проверка значимости выполняется по-своему, а полученные результаты оказываются наилучшими в том случае, когда используемые вами преобразования приводят к линейной модели множественной регрессии для ваших данных.

Прогнозирование Y весьма существенно зависит от того, подвергалась ли Y преобразованиям. Если переменная Y не подвергалась преобразованиям, уравнение регрессии прогнозирует Y непосредственно. Достаточно для каждой X_i взять соответствующим образом преобразованные значение, умножить его на коэффициент регрессии b_i, сложить все эти произведения, добавить a — и вы получаете прогнозируемое значение Y.

Преобразование переменной Y с помощью натурального логарифма может привести к коррекции имеющейся до преобразования у переменной Y асимметрии. Использование в уравнении регрессии надлежащим образом преобразованных значений X-переменных дает прогноз log Y. Новая процедура прогнозирования исходной (непреобразованной) переменной Y, представленная в приведенной выше таблице, делает две вещи. Во-первых, путем экспоненцирования прогнозированное значение log Y преобразуется к исходным единицам Y. Во-вторых, коррекция асимметрии (основанная на S_e) увеличивает это значение, отражая тот факт, что среднее значение больше, чем медиана или мода для этого вида асимметричного распределения.

Пример. *Рекламные объявления в журналах: использование преобразования и интерпретация*

В табл. 12.3.2 представлены результаты множественной регрессии для нашего примера с рекламными объявлениями в журналах после преобразования с помощью логарифма тарифа на размещение рекламы в журналах, величины читательской аудитории и медианы дохода. Теперь мы имеем дело с логарифмом тарифа на размещение рекламы в журналах (новая переменная Y), который объясняется логарифмом величины читательской аудитории (новая переменная X_1), процентом читателей-мужчин (переменная X_2) и логарифмом медианы дохода (новая переменная X_3). Попытаемся интерпретировать полученные результаты.

Значение R^2=80,5% интерпретируется обычным образом, как и в терминах исходных (непреобразованных) переменных. Это значение свидетельствует о том, что 80,5% изменчивости величины тарифа на размещение рекламы в различных журналах могут объясняться известными для каждого журнала значениями размера читательской аудитории, процента читателей-мужчин и медианы дохода читателей.[26]

---

[23] Взаимосвязи между переменными легче интерпретировать, если для преобразования Y использовать *натуральный логарифм* (по основанию e = 2,71828..., иногда, в отличие от логарифма по основанию 10, в таком случае используют обозначение "ln").

[24] Здесь мы предполагаем, что каждое преобразование является "приемлемым" — в том смысле, что оно не изменяет взаимной упорядоченности наблюдений и является относительно "гладкой" функцией.

[25] Считается, что случайная переменная имеет *логнормальное распределение*, если распределение ее логарифма является нормальным. Можно указать несколько превосходных технических описаний этого распределения, в том числе книги Johnson N. L. and Kotz S. *Continuous Univariate Distributions* (New York: Wiley, 1970), Chapter 14; и Aitchison J. and Brown J. A. C. *The Lognormal Distribution* (London: Cambridge University Press, 1957). Логнормальное распределение также имеет большое значение в теории ценообразования финансовых опций.

[26] Конкретной мерой изменчивости, используемой в нашем случае, является дисперсия измеренного на логарифмической шкале тарифа на размещение рекламы в журналах, объясняемая

Смысл R^2 не меняется, независимо от того, проводились ли логарифмические преобразования, но детали несколько разнятся.

Стандартная ошибка оценки, $S_e = 0{,}2603$, получает новую интерпретацию. Чтобы придать смысл этому числу (которое буквально означает типичную величину ошибок прогнозирования на логарифмической шкале), воспользуемся следующим уравнением:

$$\sqrt{2{,}71828^{(S_e^2)} - 1} = \sqrt{2{,}71828^{(0{,}2603^2)} - 1} = \sqrt{2{,}71828^{0{,}06776} - 1} = \sqrt{1{,}0701 - 1} = 0{,}265, \text{ или } 26{,}5\%.$$

Таблица 12.3.1. Интерпретация множественной регрессии с использованием преобразования

	Если переменная *Y* не преобразовывалась	Если использовалось натуральное логарифмирование *Y*
R^2	*Обычная интерпретация.* Процент изменчивости *Y*, объясняемый *X*-переменными (возможно, преобразованными)	*Обычная интерпретация.* Процент изменчивости переменной *Y* (преобразованной), который объясняется *X*-переменными (возможно, преобразованными)
Se	*Обычная интерпретация.* Приблизительная величина ошибок прогнозирования *Y*	*Новая интерпретация. Коэффициент вариации* ошибок прогнозирования *Y*, который задается выражением* $\sqrt{2{,}71828^{(S_e^2)} - 1}$
b_i	*Обычная интерпретация.* Ожидаемое влияние единичного изменения X_i (возможно, преобразованного) на *Y*, когда все остальное не меняется	*Аналогичная интерпретация.* Ожидаемое влияние единичного изменения X_i (возможно, преобразованного) на log *Y*. Если и X_i преобразовано с помощью логарифма, то b_i также называются эластичностью *Y* по отношению к X_i: ожидаемое влияние (в процентных единицах *Y*) изменения X_i на 1%, когда все остальное не меняется
Тест значимости для b_i	*Обычная интерпретация.* Влияет ли X_i на *Y*, когда все другие *X*-переменные остаются неизменными?	*Обычная интерпретация.* Влияет ли X_i на *Y*, когда все другие *X*-переменные остаются неизменными?
Прогнозирование *Y*	*Обычная процедура.* Использовать уравнение регрессии для прогнозирования *Y* на основании *X*-переменных, сначала преобразовав *X*-переменные	*Новая процедура.* Начните с использования уравнения регрессии для прогнозирования log *Y* на основании *X*-переменных, сначала преобразовав *X*-переменные. Затем найдите прогнозируемое значение Y, воспользовавшись следующим выражением†: $2{,}71828^{[(1/2)S_e^2 + \text{Прогнозируемое значение для log } Y]}$

*Предостережение. Этот коэффициент вариации может оказаться недостоверным в случае, если он принимает значения, существенно превышающие 1 (или 100%), поскольку в таких случаях очень большая асимметрия затрудняет оценивание средних и стандартных отклонений.

†Это выражение прогнозирует ожидаемое (т.е. среднее) значение *Y* для заданных значений *X*-переменных. Если же требуется прогнозировать медианное значение *Y*, можно воспользоваться следующей, более простой формулой:

$$2{,}71828^{(\text{Прогнозируемое значение для log } Y)}.$$

в рамках модели множественной линейной регрессии с помощью логарифма величины читательской аудитории, процента читателей-мужчин и логарифма среднего дохода.

Таблица 12.3.2. Распечатка результатов множественной регрессии после логарифмического преобразования тарифа на размещение рекламы в журналах, величины читательской аудитории и среднего дохода

Уравнение регрессии имеет следующий вид:

log Тариф на размещение рекламы = –3,44 + 0,578 (log Аудитория) – 0,00163 (процент мужчин) + 0,890 (log Доход).

Независимая переменная	Коэффициент	Стандартное отклонение	t	p
Константа	–3,441	2,011	–1,71	0,093
log Аудитория	0,57847	0,04023	14,38	0,000
Процент мужчин	–0,001635	0,001613	–1,01	0,316
log Доход	0,8897	0,1793	4,96	0,000

S = 0,2603 R-квадрат = 80,5% R-квадрат (коррект.) = 79,4%

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	3	14,2737	4,7579	70,22	0,000
Ошибка остатка	51	3,4557	0,0678		
Итого	54	17,7294			

Источник	DF	Seq SS
log Аудитория	1	12,4115
Процент мужчин	1	0,1945
log Доход	1	1,6677

Необычные наблюдения

Наблюдения	log Аудитория	log Тариф на рекламу	Соответствие	Стандартное отклонение соответствия	Остаток	Стандартизованный остаток
9	8,3	10,2421	10,8394	0,0705	–0,5973	–2,38R
48	8,2	9,7468	9,8707	0,1889	–0,1239	–0,69X
54	8,1	11,3794	10,7636	0,0707	0,6158	2,46R
55	8,0	11,2019	10,6940	0,0629	0,5079	2,01R

R обозначает наблюдение со значительным стандартизованным остатком.

X обозначает наблюдение, на которое X-значение обеспечивает существенное влияние.

Это свидетельствует о том, что ваша ошибка прогнозирования в типичном случае составляет 26,5% от прогнозируемого значения. Если, например, ваш прогнозируемый тариф на размещение рекламы в журнале равен $100 000, вариация составляет 26,5% от этого значения, или $26 500, что дает стандартную ошибку оценки для тарифа на размещение рекламы в журналах, которое вполне применимо к такого рода очень большим журналам. Если же ваш прогнозируемый тариф на размещение рекламы в журналах равен $20 000, взяв 26,5% от этого значения, получим $5 300 как соответствующую стандартную ошибку для подобного рода небольших журналов. В том, что стандартная ошибка оценки должна зави-

сеть от масштаба журнала, есть определенный смысл, поскольку большие журналы имеют гораздо больше возможностей для изменчивости, чем небольшие.

Коэффициент регрессии $b_1 = 0{,}578$ (для логарифма величины читательской аудитории) представляет собой эластичность, поскольку преобразование с помощью натуральных логарифмов использовалось и для Y. Таким образом, увеличение читательской аудитории на 1% позволяет нам рассчитывать на увеличение тарифа на размещение рекламы в журнале на 0,578%. Это указывает на наличие эффекта уменьшенного отклика, в результате которого увеличение читательской аудитории на 1% приводит к несколько меньшему (т.е. меньше, чем на 1%) увеличению тарифа на размещение рекламы. У вас может возникнуть вопрос, действительно ли это уменьшение является значимым или коэффициент $b_1 = 0{,}578$, по существу, равен 1 — если не принимать во внимание действие случайного фактора. Ответ заключается в том, что указанное заданное значение 1 находится за пределами доверительного интервала для b_1 (который расположен между 0,498 и 0,659), а это свидетельствует о значимом уменьшении. К этому выводу можно было бы прийти и другим путем — вычислив t-статистику: $t = (0{,}578 - 1)/0.0402 = -10{,}5$.

Оказывает ли величина читательской аудитории значимое влияние на величину рекламного тарифа, если процент читателей-мужчин и средний доход остаются неизменными? Ответ на этот вопрос является положительным, о чем свидетельствует обычный t-тест значимости b_1 в данной множественной регрессии. К этому выводу можно прийти на основе p-значения (в табл. 12.3.2 это значение равняется 0,000 для независимой переменной "**log** Аудитория").

И наконец, давайте определим прогнозируемое значение Y для журнала *Audubon*. Это значение будет несколько отличаться от прогнозируемого значения, вычисленного намного раньше в этой главе; к тому же оно оказывается несколько лучшим, так как до преобразования исследуемые данные не соответствовали модели линейной множественной регрессии. Прогнозирование Y выполняется в два этапа: сначала мы прогнозируем log Y непосредственно из уравнения регрессии, а затем используем S_e для получения прогнозируемого значения.

Журнал *Audubon* характеризуется следующими значениями: $X_1 = 1\,645$ (т.е. читательская аудитория этого журнала равна 1,645 миллиона человек), $X_2 = 51{,}1$ (указывает на то, что среди читателей этого журнала 51,1% мужчин) и X3 = $38 787 (указывает медиану дохода семьи читателей этого журнала). Преобразуя в уравнении регрессии величины читательской аудитории и среднего дохода с помощью логарифма, находим прогнозируемое значение для log (тариф на размещение рекламы в журналах) для журнала *Audubon*.

Прогнозируемое значение log (тариф на размещение рекламы в журналах) =
$-3{,}441 + 0{,}57847 \times \log(\text{Аудитория}) - 0{,}001635(\text{процент читателей-мужчин}) +$
$+ 0{,}8897 \times \log(\text{Доход}) = -3{,}441 + 0{,}57847 \times \log(1\,645) - 0{,}001635(51{,}1) +$
$+ 0{,}8897 \times \log(38\,787) = -3{,}441 + 0{,}57847 \times 7{,}4055 - 0{,}001635(51{,}1) +$
$+ 0{,}8897 \times 10{,}5658 = -3{,}441 + 4{,}2839 - 0{,}0835 + 9{,}4004 = 10{,}160.$

Чтобы найти прогнозируемое значение тарифа на размещение рекламы в журналах, нужно выполнить следующий этап:

прогнозируемый тариф на размещение рекламы в журналах = $e^{[(1/2)S_e^2 + \text{Прогнозируемое значение для log Y}]}$

$= 2{,}71828^{[(1/2)0{,}2603^2 + 10{,}160]} = 2{,}71828^{10{,}1939} = \$26\,739$.

Это прогнозируемое значение сравнимо с фактической величиной рекламного тарифа для этого журнала — $25 315. Нам повезло, что эти значения достаточно близки друг к другу. Соответствующая стандартная ошибка для сравнения фактического и прогнозируемого значений составляет 26,5% от $26 739, что равняется $7 086. Если вы вычислите прогнозируемую величину рекламного тарифа для других журналов, то окажется, что они, как правило, не настолько близки к фактическим значениям, как в рассмотренном нами случае. Для сравнения можно взглянуть на относительные ошибки прогнозирования для первых десяти журналов из всего перечня журналов в нашем примере: –5,6%; 11,3%; 27,7%; –23,3%; 1,9%; 0,9%; 18,8%; 8,0%; –88,0% и –21,6%. Исходя из этого величина 26,5% представляется вполне приемлемым вариантом типичной величины ошибок.

Подгонка кривой с помощью полиномиальной регрессии

Рассмотрим нелинейную *двумерную* взаимосвязь. Если диаграмма рассеяния Y в зависимости от X демонстрирует наличие нелинейной взаимосвязи, можно попытаться воспользоваться множественной регрессией, введя сначала новую X-переменную, взаимосвязь которой с переменной X также является нелинейной. Простейшим вариантом является введение переменной X^2 — квадрата исходной переменной X. Теперь вы имеете дело с *многомерной* совокупностью данных, которая характеризуется наличием трех переменных: Y, X и X^2. Когда вы прогнозируете Y на основании одной переменной X и некоторых из ее степеней (X^2, X^3 и т.д.), вы имеете дело с **полиномиальной регрессией**. Рассмотрим случай использования переменной X вместе с X^2.

В случае использования этих переменных обычное уравнение множественной регрессии, $Y = a + b_1X_1 + b_2X_2$, превращается в *квадратичный полином* — $Y = a + b_1X + b_2X^2$.[27] Такая взаимосвязь по-прежнему рассматривается как линейная, поскольку отдельные члены складываются. Точнее говоря, вы имеете дело с *линейной* взаимосвязью между Y и парой переменных (X, X^2), которые вы используете для объяснения *нелинейной* взаимосвязи между Y и X.

Начиная с этого момента вы можете просто вычислить множественную регрессию Y по двум переменным X и X^2 (таким образом, количество переменных увеличивается до $k = 2$, в то время как количество наблюдений, n, не изменяется). При этом используется вся рассмотренная ранее техника: прогнозы, остатки, R^2 и S_e, как меры качества регрессии, тесты для коэффициентов регрессии и т.д.

На рис. 12.3.1 представлены некоторые из множества кривых, которые могут порождаться квадратичными полиномами. Если ваша диаграмма рассеяния Y в зависимости от X похожа на одну из этих кривых, то введение X^2 в качестве новой переменной может быть полезным для объяснения и прогнозирования соответствующей взаимосвязи.

Пример. *Оптимизация объема производимой продукции*

Проанализируем данные из табл. 12.3.3, представляющие собой часть эксперимента для определения температуры, позволяющей получить наибольший объем продукции, выпускаемой в ходе некоторого производственного процесса. Эти данные могут оказаться чрезвычайно полезными для вашей фирмы, поскольку они свидетельствует о том, что для максимизации объема выпускаемой продукции температуру процесса следует установить около 700 градусов. Объем выпускаемой продукции заметно снижается, если температура существенно отличается от указанного значения (в ту или другую сторону).

Диаграмма рассеяния, показанная на рис. 12.3.2 с помощью линии наименьших квадратов, демонстрирует, сколь неподходящей может оказаться линейная регрессия при попытках ее использования для прогнозирования нелинейной взаимосвязи. Структуру, которая в данном случае просматривается совершенно четко, можно использовать для прогнозирования объема выпускаемой продукции на основе температуры и для определения температуры, обеспечивающей максимальный объем продукции, но прямая линия в данном случае совершенно неуместна!

[27] Слово *полином* обозначает любую сумму неотрицательных целых степеней некоторой переменной, умноженных на постоянные коэффициенты, например $3 + 5x - 4x^2 - 15x^3 + 8x^6$. Слово *квадратный* означает, что в соответствующем полиноме не может быть степеней, больших 2, например: $7 - 4x + 9x^2$ или $9 - 3x^2$. Несмотря на то что для моделирования более сложных нелинейных взаимосвязей, в принципе, можно использовать полиномы более высоких степеней, в случае степеней, превышающих 3, результаты зачастую оказываются нестабильными.

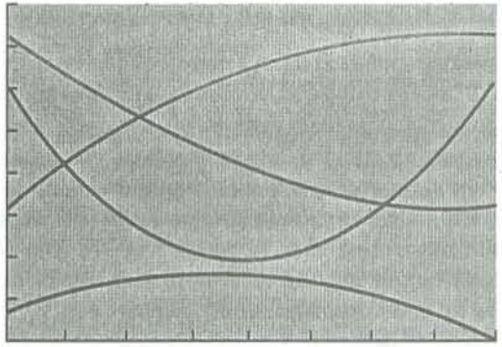


Рис. 12.3.1 Квадратичные полиномы можно использовать для моделирования самых различных нелинейных взаимосвязей. На этом рисунке представлено лишь несколько возможных вариантов. Переворачивание любой из этих кривых по горизонтали или по вертикали по-прежнему дает вам квадратичный полином

Таблица 12.3.3 Температура и объем продукции производственного процесса

Температура, *X*	Объем продукции, *Y*	Температура, *X*	Объем продукции, *Y*
600	127	750	153
625	139	775	148
650	147	800	146
675	147	825	136
700	155	850	129
725	154		

Эту проблему способна решить полиномиальная регрессия; кроме того, она даст вам надежную оценку оптимальной температуры, обеспечивающей максимальный объем выпускаемой продукции. В табл. 12.3.4 представлена многомерная совокупность данных, которая будет использоваться в этом случае. Обратите внимание, что новой является лишь последняя переменная (квадрат температуры). Ниже представлено уравнение прогнозирования, полученное методом множественной регрессии. На рис 12.3.3 представлен соответствующий график и данные.

$$\text{Объем продукции} = -712{,}10490 + 2{,}39119\ (\text{температура}) - 0{,}00165\ (\text{температура}^2).$$

Коэффициент детерминации для этой множественной регрессии, $R^2 = 0{,}969$, свидетельствует, что очень большая часть вариации объема выпускаемой продукции, а именно 96,9%, объясняется температурой и ее (температуры) квадратом. (В действительности сама по себе прямая линия объясняет менее 1%.) Стандартное отклонение оценки $S_e = 1{,}91$ указывает на то, что объем выпускаемой продукции можно прогнозировать с точностью в несколько единиц (сравните с соответствующим намного большим значением 10,23 для прямой линии).

Как проверить, действительно ли нам необходим дополнительный член (квадрат температуры)? *t*-тест для соответствующего коэффициента регрессии ($b_2 = -0{,}00165$), основанный на стандартной ошибке $S_{b_2} = 0{,}000104$ с 8 степенями свободы, указывает на очень высокую значимость этого члена уравнения. Разумеется, это было очевидно из сильной кривизны на диаграмме рассеяния. Соответствующие результаты представлены в табл. 12.3.5.

Какую температуру лучше всего использовать для оптимизации объемов выпуска продукции? Если коэффициент регрессии b_2 для квадрата переменной X является отрицательным (как в данном случае), то квадратичный полином принимает максимальное значение при $-b_1/2b_2$.[28] В нашем случае температура, обеспечивающая максимальный объем выпускаемой продукции, определяется следующим образом:

$$\text{оптимальная температура} = -b_1/2b_2 = -2{,}39119/[2(-0{,}00165)] = 724{,}6.$$

Таким образом, правильным будет установить температуру на уровне 725 градусов.

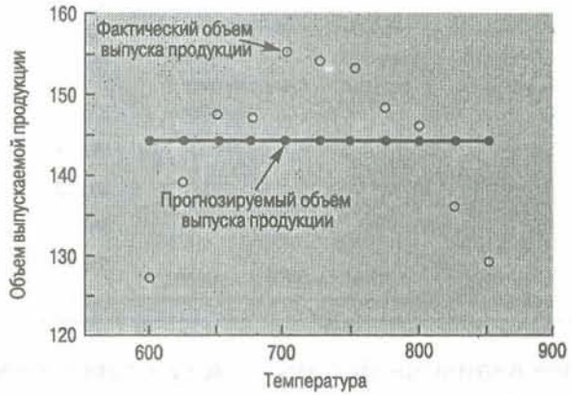


Рис. 12.3.2. Нелинейная взаимосвязь между объемом выпускаемой продукции и температурой производственного процесса чрезвычайно плохо описывается линией наименьших квадратов. Прогнозируемые значения не имеют почти ничего общего с фактическими значениями

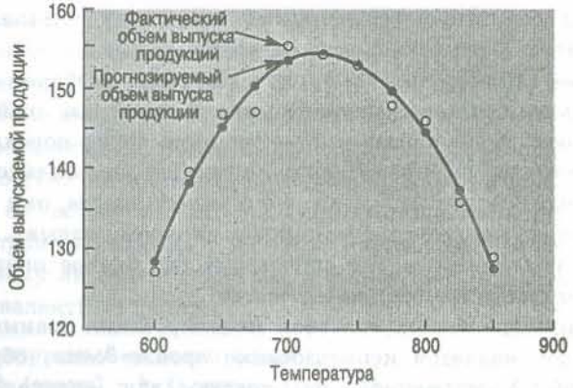


Рис. 12.3.3. Результаты регрессии с помощью квадратичного полинома хорошо объясняют объем выпускаемой продукции на основе температуры и квадрата температуры. Теперь прогнозы можно считать почти идеальными

[28] Если b_2 является положительным числом, тогда в той же точке $(-b_1/2b_2)$ полином будет иметь *минимальное* значение.

Таблица 12.3.4. Создание новой переменной (квадрат температуры) для использования полиномиальной регрессии

Объем продукции, Y	Температура, $X_1=X$	Квадрат температуры, $X_2=X^2$
127	600	360 000
139	625	390 625
147	650	422 500
147	675	455 625
155	700	490 000
154	725	525 625
153	750	562 500
148	775	600 625
146	800	640 000
136	825	680 625
129	850	722 500

Моделирование взаимодействия между двумя *X*-переменными

В модели линейной множественной регрессии каждая из X-переменных умножается на свой коэффициент регрессии; затем все эти компоненты (и константа a) складываются, обеспечивая требуемый прогноз: $a + b_1X_1 + b_2X_2 + ... + b_kX_k$. В этом выражении никак не учитывается взаимодействие между X-переменными. Говорят, что между двумя переменными наблюдается **взаимодействие**, если изменение значений обеих этих переменных приводит к ожидаемому изменению Y, которое отличается от суммы изменений Y, вызываемых изменением каждой из этих X-переменных по отдельности.

Взаимодействие наблюдается во многих системах, особенно если для успеха требуется правильное сочетание ингредиентов. Рассмотрим крайний случай. Допустим, X_1 — порох, X_2 — нагрев и Y — реакция. Фунт пороха сам по себе не представляет опасности, да и зажженная спичка не дает сильного эффекта сама по себе. Но вот если эти две переменные соединить вместе, они вступят во взаимодействие и приведут в качестве реакции к сильному взрыву. В сфере бизнеса взаимодействие проявляется в тех случаях, когда "целое оказывается больше (или меньше), чем сумма его составных частей".

Одним из распространенных способов моделирования взаимодействия в регрессионном анализе является использование *произведения*, образуемого путем умножения одной X-переменной на другую для определения новой X-переменной, которая включается — наряду с другими — в вашу множественную регрессию. Такое произведение представляет взаимодействие этих двух переменных. Более того, можно будет выполнить проверку на наличие взаимодействия, воспользовавшись t-тестом значимости коэффициента регрессии для этого терма взаимодействия.

Если вы имеете дело с каким-либо важным взаимодействием, но оно не учитывается в уравнении регрессии, ваши прогнозы окажутся весьма далекими от

Таблица 12.3.5. Результаты множественной регрессии с использованием квадрата температуры в качестве переменной для получения полиномиальной регрессии

$S = 1,907383$

$R^2 = 0,969109$

Статистический вывод на уровне 5% относительно объема выпускаемой продукции

Уравнение прогнозирования действительно объясняет значимую долю вариации объема выпускаемой продукции.

$F = 125,4877$ с 2 и 8 степенями свободы

	Влияние на объем выпускаемой продукции	95% доверительный интервал		Проверка гипотез	Стандартная ошибка коэффициента	t-статистика
Переменная	Коэффициент	От	До	Значимый?	Стандартная ошибка	t
Константа	–712,104	–837,485	–586,723	Да	54,37167	–13,0969
Температура	2,391188	2,042414	2,739963	Да	0,151246	15,80988
Температура^2	–0,00165	–0,00189	–0,00141	Да	0,000104	–15,8402

действительности. Рассмотрим, например, прогнозирование объема продаж (Y) на основе протяженности командировок (X_1, мили) и количества контактов (X_2, количество людей, с которыми были встречи) для некоторой группы коммивояжеров. Обычное уравнение регрессии, которое можно было бы использовать для прогнозирования объема продаж, $a + b_1$ (мили) + b_2 (контакты), не учитывает возможность взаимодействия "миль" и "количества контактов". Ценность дополнительной "командировочной мили" (сама по себе) оценивается как b_1 — *независимо от количества встреч с людьми при этом*. Аналогичным образом ценность дополнительной встречи (сама по себе) оценивается как b_2 — *независимо от количества потребовавшихся для этого миль командировок*.

Если вам кажется, что между "милями" и "количеством контактов" существует какое-то взаимодействие, в результате которого коммивояжеры, имеющие больше контактов с людьми, более продуктивно использовали свои "командировочные мили", то указанная модель не отражает действительности. Один из способов исправить ее состоит в том, чтобы ввести новую X-переменную, представляющую собой произведение $X_3 = X_1 \times X_2 =$ контакты × мили. Результирующая модель по-прежнему является линейной и может быть представлена в двух различных, но эквивалентных формах:

прогнозируемый объем продаж =
$= a + b_1$ (мили) + b_2 (контакты) + b_3 (контакты × мили) =
$= a + [b_1 + b_3$ (контакты)] (мили) + b_2 (контакты).

Это выражение говорит о том, что дополнительная "командировочная миля" значит для объема продаж больше, если количество контактов оказывается большим (при условии, что $b_3 > 0$). Вы можете использовать t-тест для b_3 для определения значимости этого влияния; если это влияние не значимо, дополнительную переменную X_3 можно просто отбросить и строить регрессию Y по X_1 и X_2.

Еще один способ моделирования взаимодействия в регрессионном анализе заключается в преобразовании некоторых или всех переменных. Поскольку логарифмирование преобразует умножение в сложение, мультипликативное уравнение с взаимодействием,

$$Y = AX_1^{b_1}X_2^{b_2},$$

после логарифмирования всех переменных преобразуется в следующее линейное аддитивное уравнение без взаимодействия:

$$\log Y = \log A + b_1 \log X_1 + b_2 \log X_2 = \mathrm{a} + b_1 \log X_1 + b_2 \log X_2.$$

12.4. Индикаторные переменные: прогнозирование на основе категорий

Множественная регрессия базируется на арифметике и, следовательно, требует осмысленных чисел (количественных данных). А что делать, если не все переменные являются количественными? **Индикаторная переменная**, которую также называют *фиктивной переменной* (dummy variable), — это количественная переменная, которая принимает только два значения, 0 и 1, и используется для представления качественных категориальных данных. Например, у вас может быть переменная, представляющая пол, которая равна 1 для женщин и 0 для мужчин (или наоборот, если вам так нравится больше). В анализе множественной регрессии можно использовать одну или несколько индикаторных переменных в качестве независимых (X) переменных.[29]

Если качественная X-переменная включает в точности две категории (например, мужчины/женщины, покупать/приценяться или негодный/годный), ее можно представить непосредственно как индикаторную переменную. Вы можете принять волевое решение по поводу того, какая из категорий будет соответствовать 1, а какая — 0 (база). Несмотря на то что ваш выбор на данном этапе является произвольным, необходимо помнить, что вариант кодирования влияет на интерпретацию результатов, которые вы получите впоследствии. В табл. 12.4.1 показан пример категориальной переменной, которая представляет пол каждого из респондентов (1 соответствует женщинам, 0 — мужчинам).

Если качественная X-переменная включает более двух категорий, то чтобы заменить ее, вам придется воспользоваться несколькими индикаторными переменными. Прежде всего выберите одну из категорий, которая будет служить в качестве базового значения, по отношению к которому будет измеряться влияние всех других категорий. *Не* используйте в регрессионном анализе индикаторную переменную для базовой категории, поскольку эта категория будет представлена в уравнении регрессии постоянным членом. Для каждой из всех остальных (т.е. отличных от базовой) категорий необходимо создать отдельную индикаторную переменную. Для каждой элементарной единицы (человека, фирмы или чего-нибудь другого) из выборки у вас будет не более одного значения 1 в группе индикаторных переменных; все они будут равны 0, если эта элементарная единица принадлежит к базовой категории. Помните следующее правило.

Таблица 12.4.1. Индикаторная переменная, представляющая пол человека

Категориальная переменная	Индикаторная переменная
Мужчина	0
Мужчина	0
Женщина	1
Мужчина	0
Женщина	1
Женщина	1
.	.
.	.

Правило использования индикаторных переменных

Количество индикаторных переменных, используемых во множественной регрессии для замены переменной качественного типа, должно быть *на одну меньше* количества категорий. Оставшаяся категория определяет базу. Базовая категория представляется в уравнении регрессии постоянным членом.

Какую категорию выбрать в качестве базовой? Можно выбрать ту, с которой вы хотели бы сравнивать все остальные категории.[30] Можно, например, выбрать категорию, которая встречается чаще других.

Вот пример категориальной переменной, которая представляет в выборке тип объектов, обрабатываемых отделом почтовой корреспонденции фирмы. Используется четыре категории: бизнес-конверт, большой конверт, небольшая коробка и большая коробка. Поскольку большинство почтовой корреспонденции фирмы составляет бизнес-конверты, именно этот вид корреспонденции выбран в качестве базовой категории. Эту качественную переменную (тип объекта) предполагается использовать в анализе множественной регрессии для объяснения Y = время обработки. В табл. 12.4.2 показаны три индикаторные переменные, которые необходимо создать и использовать вместе с другими X-переменными.

[29] Если ваша зависимая переменная (Y) является качественной, то ситуация оказывается намного более сложной, поскольку тогда терм ошибки, ε, в модели линейной множественной регрессии не может иметь нормального распределения. Если Y имеет два возможных значения, можно воспользоваться так называемой *логит-моделью* (*множественная логистическая регрессия*) или *пробит-моделью*. Если переменная Y может принимать более двух различных значений, вам может подойти *полиномиальная логит-модель* (multinomial logit model) или *полиномиальная пробит-модель* (multinomial probit model). Эти вопросы освещены в книге Kmenta J. *Elements of Econometrics* (New York: Macmillan, 1986), Section 11-5.

[30] А что если вам придется выполнять сравнение с несколькими категориями? Одно достаточно простое решение заключается в том, чтобы выполнить *несколько* анализов множественной регрессии — каждый с использованием своей собственной базовой категории.

Интерпретация и проверка значимости коэффициентов регрессии для индикаторных переменных

После того как категориальные X-переменные заменены на индикаторные переменные, множественную регрессию можно выполнять обычным способом. Несмотря на то что и в этом случае регрессию можно интерпретировать обычным образом, существует несколько особых способов интерпретации коэффициентов регрессии и их t-тестов в случае использования индикаторных переменных (табл. 12.4.3). Помните: если X_i является индикаторной переменной, она представляет только одну категорию исходной качественной переменной (а именно категорию, для которой она равна 1).

Таблица 12.4.2. Использование трех индикаторных переменных для представления четырех категорий, исключая бизнес-конверт как базовую категорию

	Индикаторные переменные		
Категориальная переменная: тип объекта	**Большой конверт, X_1**	**Небольшая коробка, X_2**	**Большая коробка, X_3**
Бизнес-конверт	*0*	*0*	*0*
Небольшая коробка	0	1	0
Бизнес-конверт	0	0	0
Бизнес-конверт	0	0	0
Большая коробка	0	0	1
Большой конверт	1	0	0
Бизнес-конверт	0	0	0
Большой конверт	1	0	0
Бизнес-конверт	0	0	0
.	.	.	.
.	.	.	.

Таблица 12.4.3. Интерпретация коэффициента регрессии для индикаторной переменной X_i

b_i	Коэффициент регрессии b_i представляет *среднюю разницу* значений Y в двух категориях — той, которую представляет X_i, и базовой категорией (другие X-переменные X при этом остаются неизменными). Если b_i является положительным числом, эта категория имеет *более высокое* значение оценки среднего Y, чем базовая категория. Если b_i – отрицательное число, то среднее Y для этой категории оказывается *ниже*, чем для базы (при всех прочих равных)
Проверка значимости для b_i	С точки зрения ожидаемого значения Y (другие X-переменные при этом остаются неизменными), есть ли какая-то разница (кроме случайности) между категорией, которую представляет X_i, и базовой категорией?

Пример. *Оценка влияния пола работника на уровень заработной платы (с поправкой на стаж работы)*

Руководство вашей фирмы озабочено возможностью обвинений в дискриминации сотрудников по признаку пола. Кое у кого возникают подозрения, что в вашем отделе сотрудники-мужчины зарабатывают больше, чем женщины. Краткий анализ заработной платы *24 мужчин и 26 женщин, работающих в вашем отделе*, показывает, что в среднем мужчина получает за год на $4 214 больше женщины. Более того, принимая во внимание стандартную ошибку, равную $1 032, можно утверждать, что эта разница является высоко статистически значимой ($p < 0{,}001$).[31]

Означает ли это, что дискриминация сотрудников по признаку пола действительно имеет место в вашей фирме? Вообще говоря, необязательно. Указанные статистические результаты действительно суммируют заработные платы двух категорий работников и сравнивают полученную разницу с тем, что можно было бы воспринимать как результат действия фактора случайности. С точки зрения статистики можно было бы сделать вывод о том, что различия работников по полу в определенной мере (значительно выходящей за рамки простой случайности) влияют на уровень их заработной платы. Однако статистика ничего не говорит по поводу возможных причин этой разницы в заработной плате. Несмотря на то что в вашей фирме действительно может *быть дискриминация (более или менее значительная)* по признаку пола при найме сотрудников, существуют и другие возможности объяснения указанных различий. Могут найтись даже определенные экономические обоснования более высокого уровня оплаты мужчин в конкретной ситуации.

На собрании возникает предположение, что стаж работы также следует принимать во внимание как возможное объяснение различий в уровне заработной платы. Анализ этой причины поручается вам, и вы решаете воспользоваться множественным регрессионным анализом, чтобы понять *влияние пола сотрудника на уровень заработной платы с учетом стажа работы*. Множественная регрессия является подходящей для данного случая процедурой, поскольку коэффициент регрессии всегда включает поправку на другие X-переменные. Коэффициент регрессии для индикаторной переменной, представляющей пол сотрудника, даст вам ожидаемую разницу в уровне заработной платы между мужчиной и женщиной с одинаковым рабочим стажем.

Переменными вашей множественной регрессии являются заработная плата (Y), рабочий стаж (X_1) и пол (X_2). Пол будет представлен как индикаторная переменная, для которой *женщина — 1*, а *мужчина — 0*. Соответствующая многомерная совокупность данных представлена в табл. 12.4.4.

Здесь показаны результаты исследования данных. Диаграмма рассеяния (заработная плата как функция от стажа работы), представленная на рис. 12.4.1, свидетельствует о сильной взаимосвязи (корреляция $r = 0{,}803$) между этими переменными. Сотрудники с большим рабочим стажем, как правило, получают большую заработную плату. Здесь также просматривается определенная тенденция к нелинейности (возможно, сказывается "эффект насыщения" или содержится указание на "убывающее преимущество", когда каждый дополнительный год рабочего стажа "весит" все меньше по мере накопления все большего опыта). В любом случае можно рассчитывать на то, что стаж работы будет объяснять значительную долю вариации заработной платы.

Диаграмма рассеяния значений заработной платы в зависимости от пола, показанная на рис. 12.4.2, подтверждает тот факт, что труд мужчин, вообще говоря, оплачивается выше. Однако выводы из этих данных можно сделать намного проще, если представить их в виде блочной диаграммы с двумя прямоугольными блоками — по одному для каждого пола, — как это показано на рис. 12.4.3. Можно наблюдать отчетливую взаимосвязь между полом и уровнем заработной платы, причем труд мужчины оплачивается в среднем выше, чем труд женщины. Несмотря на то что эти два прямоугольника частично перекрываются, *разница в среднем уровне заработной платы является очень высоко значимой* (если воспользоваться t-тестом для двух независимых выборок из главы 10).

Взаимосвязь между полом и стажем работы, представленная на рис. 12.4.4, свидетельствует о том, что в среднем у мужчин стаж работы больше, чем у женщин. Нижняя часть прямоугольника для женщин показывает, что 25% женщин имеют очень незначительный стаж работы или вообще не имеют его.

[31] Это стандартная ошибка разности для двух независимых выборок. См. главу 10.

Таблица 12.4.4. Заработная плата, стаж работы и пол сотрудников

	Заработная плата (дол.), Y	Стаж работы (г.), X_1	Пол (1 — женщины, 0 — мужчины), X_2
	39 700	16	0
	28 500	2	1
	30 650	2	1
	31 000	3	1
	33 700	25	0
	33 250	15	0
	35 050	16	1
	22 800	0	1
	36 300	33	0
	35 600	29	1
	32 350	3	1
	31 800	16	0
	26 900	0	1
	37 250	19	0
	30 450	1	1
	31 350	2	1
	38 200	32	0
	38 200	21	1
	28 950	0	1
	33 950	34	0
	34 100	8	1
	32 900	11	1
	30 150	5	1
	30 800	1	0
	31 300	11	1
	33 550	18	1
	37 750	44	0
	31 350	2	1
	27 350	0	1
	35 700	19	1
	32 250	7	0
	25 200	0	1
	35 900	15	1
	36 700	14	0

	Заработная плата (дол.), Y	Стаж работы (г.), X_1	Пол (1 — женщины, 0 — мужчины), X_2
	32 050	4	1
	38 050	33	0
	36 100	19	0
	35 200	20	1
	34 800	24	0
	26 550	3	0
	26 550	0	1
	32 750	17	0
	39 200	19	0
	30 450	0	1
	38 800	21	0
	41 000	31	0
	29 900	6	0
	40 400	35	0
	37 400	20	0
	35 500	23	0
Среднее значение	33 313	13,98	52,0%
Стандартное отклонение	4 188	11,87	

Размер выборки: $n = 50$.

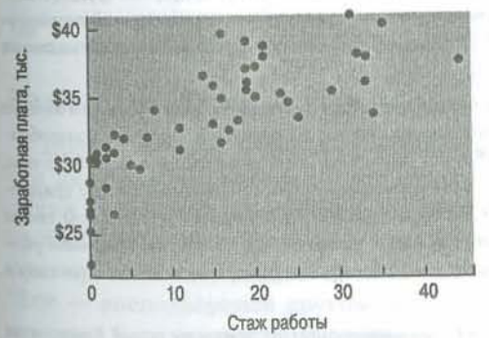


Рис. 12.4.1. Диаграмма рассеяния значений заработной платы в зависимости от рабочего стажа свидетельствует о наличии сильной взаимосвязи нарастающего типа. Труд более опытных работников оплачивается соответствующим образом

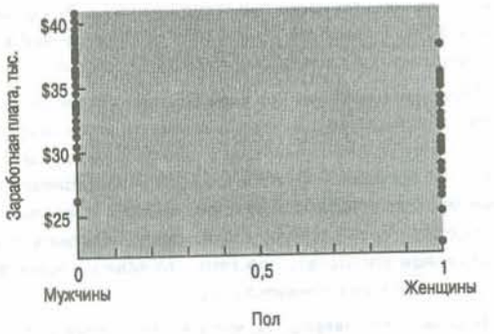


Рис. 12.4.2. Диаграмму рассеяния значений заработной платы в зависимости от пола трудно интерпретировать, поскольку пол является индикаторной переменной. В этом случае лучше воспользоваться блочной диаграммой из прямоугольников, показанной на следующем рисунке

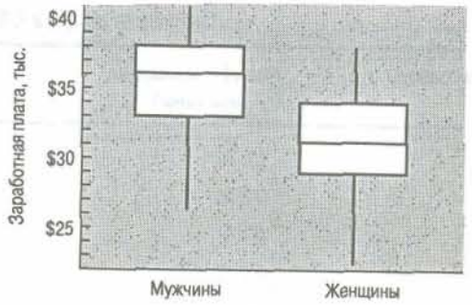


Рис. 12.4.3. Блочная диаграмма для заработной платы (по одному прямоугольнику для каждого пола) облегчает исследование взаимосвязи между полом и заработной платой. Труд мужчин в среднем оплачивается выше, чем труд женщин, хотя уровни их заработной платы в значительной мере перекрываются

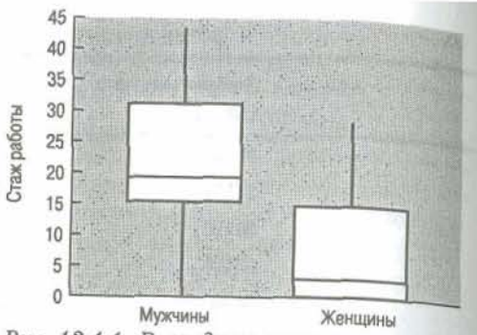


Рис. 12.4.4. В среднем мужчины имеют больший стаж работы, чем женщины. На этой блочной диаграмме представлена взаимосвязь между полом и стажем работы

Итак, какой вывод можно сделать? Наблюдается сильная взаимосвязь между всеми парами переменных. Дополнительный стаж работы компенсируется повышенной заработной платой, и у женщин отмечается более низкая заработная плата и меньший рабочий стаж.

Невыясненным остается один важный вопрос: когда вы *делаете поправку* на стаж работы (чтобы иметь возможность сравнивать заработную плату мужчин и женщин с одинаковым стажем работы), существуют ли обусловленные полом различия в заработной плате? Такой информации на диаграммах нет, поскольку вопрос включает все три переменные одновременно. Ответ на этот вопрос можно получить с помощью множественной регрессии. Соответствующие результаты представлены в табл. 12.4.5.

Коэффициент регрессии для пола, –488,08, указывает, что ожидаемая разница в уровне заработной платы между мужчиной и женщиной с одинаковым рабочим стажем равна $488,08, причем труд женщины оплачивается *ниже*, чем труд мужчины. Это объясняется тем, что увеличение индикаторной переменной X_2 на 1 приводит к переходу от 0 (мужчина) к 1 (женщина), результатом чего является ожидаемое отрицательное изменение (–$488,08) заработной платы.

Обратите внимание, что коэффициент регрессии для пола не является значимым. Более того, он весьма далек от значимости! *t*-тест значимости этого коэффициента направлен на выявление разницы в заработной плате мужчин и женщин с одинаковым стажем работы. Полученный результат свидетельствует о том, что — с поправкой на рабочий стаж — не наблюдается сколько-нибудь значимой разницы между средними уровнями заработной платы мужчин и женщин. Очевидные различия в уровнях заработной платы представителями разных полов можно объяснить различиями в их стаже работы. Итак, вы получили убедительное доказательство того, что если в вашей фирме и есть дискриминация, она основана на стаже работы, но не на признаке пола.

Но можно ли утверждать, что в вашей фирме вообще нет дискриминации по признаку пола? Вряд ли на этот вопрос можно ответить положительно. Можно лишь утверждать, что у вас нет доказательств такой дискриминации. Поскольку принятие нулевой гипотезы (обнаружено "отсутствие значимости") приводит к слабому выводу (этот вопрос обсуждался в главе 10), отсутствие дискриминации доказать сложно.

Показывает ли этот анализ, что в обществе в целом отсутствует дискриминация по признаку пола? Нет, потому что приведенные данные относятся лишь к одному отделу одной фирмы и не являются репрезентативными для общества в целом.

Однако не объясняется ли меньший стаж работы женщин дискриминацией по признаку пола, бытовавшей в нашем обществе в прошлом? Это вполне возможно, однако выполненный нами статистический анализ не дает никаких оснований для подобных утверждений. Данные, которыми мы оперировали в настоящем примере, не содержат информации ни о каких возможных причинах очевидной дискриминации в заработной плате, кроме рабочего стажа сотрудников.

Таблица 12.4.5. Результаты множественной регрессии для заработной платы, стажа работы и пола сотрудников

Уравнение регрессии имеет следующий вид:

Заработная плата

= 29776 + 271,15 * стаж работы – 488,08 * пол

Стандартная ошибка оценки,

S = 2 538,76,

указывает типичную величину ошибок прогнозирования для этой совокупности данных.

Значение R-квадрат,

R^2 = 64,7%,

указывает, какая часть дисперсии заработной платы объясняется данной регрессионной моделью.

Статистический вывод на уровне 5% относительно заработной платы

Уравнение прогнозирования действительно объясняет значимую долю вариации заработной платы.

F = 43,1572 с 2 и 47 степенями свободы

	Влияние на заработную плату	95% доверительный интервал		Проверка гипотез	Стандартная ошибка коэффициента	t-статистика
Переменная	Коэффициент	От	До	Значимый?	Стандартная ошибка	t
Константа	29776	27867	31685	Да	948,86	31,38
Стаж работы	271,15	195,46	346,84	Да	37,63	7,21
Пол	–488,08	-2269,06	1292,90	Нет	885,29	–0,55

Раздельные регрессии

Другой подход к анализу множественной регрессии многомерной совокупности данных, которая включает качественную переменную, заключается в разделении этой совокупности данных на категории с последующим выполнением отдельной множественной регрессии для каждой категории данных. Можно, например, выполнить два анализа: один для мужчин, а другой для женщин. Или — воспользуемся другим примером —отдельно проанализировать тепловые, ядерные и гидроэлектростанции.

Использование индикаторных переменных — лишь один шаг в направлении раздельных регрессий. Пользуясь индикаторными переменными, вы, в сущности, получаете отдельный для каждой категории постоянный член, но одни и те же значения для всех коэффициентов регрессии. Используя раздельные регрессии, вы получаете разные постоянные члены и разные коэффициенты регрессии для каждой категории.

12.5. Дополнительный материал

Резюме

Прогнозирование одной переменной Y на основании двух или нескольких X-переменных называется **множественной регрессией**. Целями множественной регрессии являются: (1) описание и понимание соответствующей взаимосвязи, (2) прогнозирование (предсказание) нового наблюдения, (3) регулирование и управление процессом.

Сдвиг, или **постоянный член**, a, определяет прогнозируемое значение Y при условии, что *все* X-переменные равны 0. **Коэффициент регрессии** b_j для j-й X-переменной определяет влияние переменной X_j на Y с учетом поправок на другие X-переменные; b_j указывает, какое ожидается изменение Y, когда не изменяются все X-переменные, за исключением переменной X_j, которая увеличивается на одну единицу. Взятые вместе, эти коэффициенты регрессии составляют **уравнение прогнозирования**, или **уравнение регрессии**, (прогнозируемое значение Y) $= a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$, которое можно использовать для прогнозирования или управления. Эти коэффициенты (a, b_1, b_2, ... , b_k) обычно вычисляют методом *наименьших квадратов*, который минимизирует сумму квадратов ошибок прогнозирования. **Ошибки прогнозирования**, или остатки, определяются выражением Y – (прогнозируемое значение Y).

Существуют два способа определения качества регрессионного анализа. **Стандартная ошибка оценки**, S_e, указывает приблизительную величину ошибок прогнозирования. **Коэффициент детерминации**, R^2, указывает, какой процент вариации Y объясняется (или представляется) X-переменными.

Статистический вывод начинается с проверки общей гипотезы, которую называют **F-тестом** (F-test). Цель F-теста заключается в том, чтобы выяснить, объясняют ли X-переменные значимую долю вариации Y. Если ваша регрессия *не* является значимой, говорить больше не о чем. Если же регрессия оказывается значимой, можно приступать к статистическому выводу, используя **t-тесты для отдельных коэффициентов регрессии**. Доверительные интервалы и проверки гипотез для отдельных коэффициентов регрессии основываются на соответствующих им стандартных ошибках, S_{b_1}, S_{b_2}, ... , S_{b_k}. При этом используют критическое значение из t-таблицы для $n - k - 1$ степеней свободы.

Статистический вывод базируется на **модели множественной линейной регрессии**, в соответствии с которой наблюдаемое значение Y равно взаимосвязи в генеральной совокупности плюс независимые случайные ошибки, которые имеют нормальное распределение:

$$Y = (\alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_kX_k) + \varepsilon$$
$$= (\text{взаимосвязь в генеральной совокупности}) + \text{случайность},$$

где ε характеризуется нормальным распределением со средним значением 0 и постоянным стандартным отклонением σ, причем эта случайность является независимой для каждого из наблюдений. Для всех параметров генеральной совокупности (α, β_1, β_2, ..., β_k, σ) имеются соответствующие выборочные оценки (a, b_1, b_2, ... , b_k, S_e).

В F-тесте используются следующие статистические гипотезы:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$;

H_1: по крайней мере один из коэффициентов регрессии β_1, β_2, ... , $\beta_k \neq 0$.

Результат F-теста определяется следующим образом.

Если значение R^2 оказывается *меньшим*, чем критическое значение в таблице, то соответствующая модель является *незначимой* (следует принять нулевую гипотезу о том, что X-переменные *не* позволяют прогнозировать Y).

Если значение R^2 оказывается *большим*, чем критическое значение в таблице, то соответствующая модель является *значимой* (следует отвергнуть нулевую гипотезу и принять альтернативную гипотезу о том, что X-переменные *действительно* позволяют прогнозировать Y).

Доверительный интервал для отдельного коэффициента регрессии, β_j, определяется следующим образом:

$$\text{от } b_j - tS_{b_j} \text{ до } b_j + tS_{b_j}$$

где t берется из t-таблицы для $n - k - 1$ степеней свободы. Гипотезы для t-теста j-го коэффициента регрессии имеют следующий вид:

$$H_0: \beta_j = 0;$$
$$H_1: \beta_j \neq 0.$$

Существуют два подхода к решению трудной проблемы — принятию решения о том, какие из X-переменных вносят наибольший вклад в уравнение регрессии. **Стандартизованный коэффициент регрессии**, $b_iS_{X_i} / S_Y$, представляет собой ожидаемое изменение Y, вызванное изменением X_i и измеренное в единицах стандартных отклонений Y на стандартное отклонение X_i, когда все другие X-переменные не изменяются. Если вы не хотите делать поправку на все другие X-переменные (удерживая их без изменения), можно вместо этого сравнивать абсолютные значения коэффициентов корреляции Y с каждым из X.

Существует несколько потенциальных проблем, связанных с анализом множественной регрессии.

1. Проблема **мультиколлинеарности** возникает в тех случаях, когда некоторые из ваших объясняющих переменных (X) оказываются слишком близки между собой. Отдельные коэффициенты регрессии при этом оцениваются плохо, поскольку нет достаточной информации, чтобы решить, *какая* (или какие) из X-переменных собственно объясняют Y. Необходимо исключить из рассмотрения какие-то из переменных или переопределить какие-то из переменных (возможно, используя деление одних переменных на другие), что позволило бы увеличить различие между переменными.
2. Проблема **выбора переменных** возникает в тех случаях, когда приходится иметь дело с пространным перечнем потенциально полезных независимых X-переменных и необходимо решить, какие из этих переменных следует включать в уравнение регрессии. Использование слишком большого количества X-переменных приведет к снижению качества полученных результатов, поскольку информация будет понапрасну расходоваться на оцени-

вание ненужных параметров. Если же вы отбросите одну или несколько важных X-переменных, то качество ваших прогнозов также снизится, поскольку вы проигнорируете полезную информацию. Одно из возможных решений состоит в том, чтобы включить только те переменные, необходимость которых не вызывает сомнений, воспользовавшись для этого списком, предварительно упорядоченным в соответствии с приоритетами. Другое решение заключается в том, чтобы воспользоваться одной из автоматических процедур, таких как, например, *все подмножества* или *пошаговая регрессия*.

3. Проблема **неправильного выбора модели** включает множество различных потенциальных несоответствий между вашей конкретной задачей и моделью линейной множественной регрессии. Анализируя данные, можно выявить некоторые потенциальные проблемы, связанные с нелинейностью, неравной изменчивостью и наличием резко отклоняющихся значений. Однако даже наличие подобных проблем еще ни о чем не говорит. Несмотря на то что гистограммы некоторых переменных могут быть сильно скошенными, а некоторые диаграммы рассеяния могут быть нелинейными, модель линейной множественной регрессии и в этих случаях может быть вполне применима. Так называемая *диагностическая диаграмма* помогает понять, действительно ли обнаруженная проблема является настолько серьезной, что требует решения. Еще одна существенная проблема возникает в случае, когда приходится иметь дело с *временными рядами*. В подобной ситуации можно применять множественный регрессионный анализ, используя для каждой переменной вместо исходных значений *процентные изменения* значения этой переменной между различными периодами времени.

Диагностическая диаграмма для множественной регрессии представляет собой диаграмму рассеяния значений ошибок прогнозирования (остатков) в зависимости от прогнозируемых значений; она позволяет выяснить, действительно ли есть такие проблемы в данных, которые требуют решения. Вмешательство рекомендуется лишь в тех случаях, когда диагностическая диаграмма ясно и определенно демонстрирует наличие проблемы.

Существуют три способа решения проблемы нелинейности и/или неравной изменчивости: (1) преобразовать некоторые или все переменные; (2) ввести новую переменную или (3) воспользоваться нелинейной регрессией. Если вы выполняете преобразование, то каждую группу переменных, которые измеряются в одних и тех же базовых единицах, лучше преобразовывать одинаковым способом. Если вы преобразовываете лишь некоторые из X-переменных, но не преобразовываете Y, тогда интерпретация результатов анализа множественной регрессии в основном не меняется. Если же вы используете натуральный логарифм Y, тогда интерпретация R^2 и тестов на значимость для отдельных коэффициентов регрессии также остается неизменной, отдельные коэффициенты регрессии имеют похожую интерпретацию, а S_e нуждается в новой интерпретации.

Эластичность Y по отношению к X_i представляет собой ожидаемое *процентное* изменение Y, связанное с увеличением на 1% переменной X_i (при этом другие X-переменные остаются неизменными); эластичность оценивается с помощью

коэффициента регрессии из уравнения, в котором применяются натуральные логарифмы и для Y, и для X_i.

Еще одним способом решения проблемы нелинейности является использование **полиномиальной регрессии** для прогнозирования Y на основании единственной переменной X вместе с какими-то из ее степеней (X^2, X^3 и т.д.).

Говорят, что между двумя переменными наблюдается **взаимодействие**, если изменение в обеих этих переменных приводит к ожидаемому изменению в Y, которое отличается от суммы изменений в Y, вызываемых изменением каждой из этих X-переменных по отдельности. Взаимодействие зачастую моделируется в регрессионном анализе с помощью *произведения*, образуемого путем умножения одной X-переменной на другую для создания новой X-переменной, которая включается — наряду с другими — в множественную регрессию. Взаимодействие также можно зачастую моделировать, используя преобразования некоторых или всех переменных.

Индикаторная переменная — которую также называют *фиктивной переменной* — это количественная переменная, принимающая лишь два возможных значения (0 или 1); такая переменная используется в качестве независимой (объясняющей) X-переменной для представления качественных категориальных данных. Количество индикаторных переменных во множественной регрессии для замены качественной переменной должно быть *на единицу меньше* количества категорий. Оставшаяся категория определяет *базу*. Базовая категория представляется в результирующем уравнении регрессии постоянным членом.

Вместо использования индикаторных переменных можно находить раздельные уравнения регрессии для каждой из категорий. Это приводит к более гибкой модели с различными коэффициентами регрессии для каждой из X-переменных по каждой категории.

Основные термины

- Множественная регрессия (multiple regression), 611
- Сдвиг (intercept), или постоянный член (constant term), 614
- Коэффициент регрессии (regression coefficient), 614
- Уравнение прогнозирования (prediction equation), или уравнение регрессии (regression equation), 614
- Ошибки прогнозирования (prediction errors), или остатки (residuals), 614
- Стандартная ошибка оценки (standard error of estimate), 614
- Коэффициент детерминации (coefficient of determination), 614
- F-тест (F-test), 614
- t-тесты для отдельных коэффициентов регрессии (t-tests for individual regression coefficients), 614
- Модель множественной линейной регрессии (multiple regression linear model), 626
- Стандартизованный коэффициент регрессии (standardized regression coefficient), 645

- Мультиколлинеарность (multicollinearity), 649
- Выбор переменных (variable selection), 649
- Диагностическая диаграмма (diagnostic plot), 662
- Эластичность (elasticity), 674
- Полиномиальная регрессия (polynomial regression), 679
- Взаимодействие (interaction), 682
- Индикаторные переменные (indicator variable), 684

Контрольные вопросы

1. Ответьте на следующие вопросы, касающиеся множественной регрессии.
   а) Какие три цели множественной регрессии вы можете указать?
   б) Какого рода данные требуются для множественной регрессии?
2. Ответьте на следующие вопросы, касающиеся уравнения регрессии.
   а) Для чего используется это уравнение?
   б) Откуда берется уравнение регрессии?
   в) Как интерпретируется постоянный член уравнения регрессии?
   г) Как интерпретируется коэффициент регрессии?
3. Опишите два критерия, свидетельствующие о качестве анализа множественной регрессии.
4. а) О чем свидетельствует результат F-теста?
   б) Запишите две гипотезы F-теста.
   в) Какое значение R^2 — высокое или низкое — требуется, чтобы F-тест оказался значимым? Почему?
5. а) Что представляет собой t-тест для отдельного коэффициента регрессии?
   б) Каким образом такой тест учитывает другие X-переменные?
   в) Если F-тест незначим, можно ли продолжать анализ и тестировать отдельные коэффициенты регрессии?
6. а) Как вычисляются стандартизованные коэффициенты регрессии?
   б) Для чего они используются?
   в) В каких единицах они измеряются?
7. а) Что такое мультиколлинеарность?
   б) В чем заключается искажающее влияние высокой мультиколлинеарности?
   в) Каким образом умеренная мультиколлинеарность может сделать ваш F-тест значимым даже в том случае, когда ни один из ваших t-тестов не является значимым?
   г) Каким образом решается проблема мультиколлинеарности?
8. а) Если вы стремитесь получить наилучшие прогнозы, почему бы не включить в число X-переменных все потенциально полезные переменные?

   б) Каким образом упорядоченный по приоритетам список переменных может помочь в решении проблемы выбора переменных?
   в) Кратко опишите два автоматических метода выбора переменных.
9. а) Что такое модель линейной множественной регрессии?
   б) Перечислите три случая, когда модель линейной множественной регрессии неприменима.
   в) Какая диаграмма рассеяния может помочь вам в выявлении проблем с моделью линейной множественной регрессии?
10. а) Назовите оси диагностической диаграммы.
    б) Чем полезно отсутствие структуры в диагностической диаграмме?
11. Почему все переменные, измеряемые в одних и тех же базовых единицах, необходимо преобразовывать одинаково?
12. а) Что такое эластичность?
    б) При каких обстоятельствах коэффициент регрессии указывает эластичность Y по отношению к X_i?
13. Как полиномиальная регрессия помогает справиться с нелинейностью?
14. а) Что такое взаимодействие?
    б) Что можно сделать, чтобы включить члены взаимодействия в уравнение регрессии?
15. а) Какие переменные следует создать, чтобы включить информацию о *категориальной* переменной, которая находится среди X-переменных? Как называются эти переменные и как они создаются?
    б) Сколько индикаторных переменных необходимо создать для категориальной переменной с четырьмя категориями?
    в) Что показывает коэффициент регрессии индикаторной переменной?

Задачи



1. Вашу фирму интересуют результаты размещения рекламы в журналах как одна из составляющих оценки ее маркетинговой стратегии. По каждому рекламному объявлению вы располагаете информацией о его стоимости, объеме и количестве запросов, вызванных его появлением в журнале. В частности, вы хотите выяснить, связано ли количество потенциальных клиентов, появившихся у вашей фирмы вследствие размещения этого объявления, с его размером и затратами на его размещение. Укажите для данной задачи переменную Y, X-переменные, а также соответствующую статистику или тест.
2. Снова наступило время составления бюджета, и вы хотели бы знать ожидаемую величину отдачи (выраженную в собранной сумме) от затрат каждого дополнительного доллара на сбор неоплаченных счетов — с учетом общего количества неоплаченных счетов. Укажите для данной задачи переменную Y, X-переменные, а также соответствующую статистику или тест.

3. Чтобы обосновать судебный иск о возмещении убытков, вам нужно оценить выгоду, утраченную вашей фирмой в результате трехмесячной задержки (не по вине вашей фирмы) открытия нового лесопильного завода. Вы располагаете данными об аналогичных фирмах, касающиеся их суммарных активов, производственной мощности лесопильного завода и доходов. Что касается вашей фирмы, то вам известны ее суммарные активы и проектная производственная мощность нового лесопильного завода, но вам нужно оценить возможные доходы фирмы. Укажите для данной задачи переменную *Y*, *X*-переменные, а также соответствующую статистику или тест.

4. Производительность труда — это вопрос, заботящий всех руководителей. Вы располагаете данными о производительности труда каждого сотрудника и о других характеризующих его факторах. Вы хотели бы знать, в *какой мере эти факторы объясняют вариацию производительности* труда у разных сотрудников вашей фирмы. Укажите для данной задачи переменную *Y*, *X*-переменные, а также соответствующую статистику или тест.

5. В табл. 12.5.1 представлены данные о цене, площади холста и годе создания нескольких картин Пикассо.

   а) Составьте уравнение регрессии для прогнозирования цены картины на основании площади холста и года ее создания.

   б) Интерпретируйте коэффициент регрессии для площади холста.

   в) Интерпретируйте коэффициент регрессии для года создания картины.

   г) Какой, по вашему мнению, должна быть ожидаемая продажная цена картины, созданной в 1954 г. и имеющей площадь холста, равную 4 000 квадратных сантиметров?

   д) Какова примерно величина ошибок прогнозирования для этих картин?

   е) Какой процент вариации цен картин Пикассо можно объяснить размером холста и годом создания этих картин?

   ж) Является ли эта регрессия значимой? Представьте результаты *соответствующего теста и интерпретируйте их.*

   з) Оказывает ли площадь холста значимое влияние на цену картины (с поправкой на год ее создания)? В частности, стоят ли более крупные полотна в среднем значимо больше или значимо меньше, чем полотна меньшей площади, написанные в том же году?

   и) Оказывает ли год создания картины значимое влияние на цену картины (с поправкой на площадь ее холста)? Какой из этого можно сделать вывод о цене картин с точки зрения года их создания?

6. Рассмотрим анализ множественной регрессии для 50 штатов, который объясняет количество новых рабочих мест исходя из количества новых фирм и процента быстро развивающихся компаний. Используются следующие переменные: "новые рабочие места" (в тысячах), "новые фирмы" (фактическое количество фирм) и "процент быстрых" (в процентных единицах; таким образом, например, 3,15% представляется числом 3,15). Результаты множественной регрессии представлены в табл. 12.5.2.[32]

   а) Какая приблизительно часть вариации количества созданных новых рабочих мест между различными штатами объясняется количеством новых фирм и процентом быстро развивающихся фирм?

Таблица 12.5.1. Цена, площадь холста и год создания картин Пикассо

Цена, тыс. дол.	Площадь, см²	Год	Цена, тыс. дол.	Площадь, см²	Год
100	768	1911	360	1 141	1943
50	667	1914	150	5 520	1944
120	264	1920	65	5 334	1944
400	1 762	1921	58	1 656	1953
375	10 109	1921	65	2 948	1956
28	945	1922	95	3 510	1960
35	598	1923	210	6 500	1963
750	5 256	1923	32	1 748	1965
145	869	1932	55	3 441	1968
260	7 876	1934	80	7 176	1969
78	1 999	1940	18	6 500	1969
90	5 980	1941			

Данные получены из E. Mayer, *International Auction Records*, vol. XVII (Caine, England: Hilmarton Manor Press, 1983), p. 1056–1058.

Таблица 12.5.2. Результаты множественной регрессии для быстро развивающихся компаний

Уравнение регрессии имеет следующий вид:

новые рабочие места = –144,764 + 0,099109 * (новые фирмы) + 78,61557 * (процент быстрых)

S = 133,7854

R^2 = 81,0%

	Влияние на новые рабочие места	95% доверительный интервал		Проверка гипотез	Стандартная ошибка коэффициента	t-статистика
Переменная	Коэффициент	От	До	Значимый?	Стандартная ошибка	t
Константа	-144,76	–282,96	–6,57	Да	68,6944	–2,11
Новые фирмы	0,0991	0,0825	0,1157	Да	0,0082	12,04
Процент быстрых	78,62	20,04	137,19	Да	29,1152	2,70

[32] Анализируемые данные взяты из "Ratings the States 1987: New Jobs, New Companies, and the Climate for Growth", *INC*, October 1987, p. 77.

б) Объясняют ли переменные "Новые фирмы" и "Процент быстрых" значимую долю изменчивости переменной "Новые рабочие места"? Поясните свой ответ.

в) Найдите прогнозируемое значение переменной "Новые рабочие места" и остаток для штата Вашингтон при условии, что для этого штата новые рабочие места = 242 (тысяч), Новые фирмы = 1 741 и процент быстрых = 2,44%.

г) О чем свидетельствует коэффициент регрессии 0,0991 для "Новых фирм"?

д) Оказывает ли, по вашему мнению, процент быстро развивающихся фирм влияние на создание новых рабочих мест — с учетом поправки на количество новых компаний? Поясните свой ответ.

е) Кратко опишите (в письменном виде), что нового вы узнали о создании новых рабочих мест в результате проведенного анализа множественной регрессии.

7. Можно предположить, что цена палатки отражает различные ее характеристики. Например, можно ожидать, что большие палатки — при прочих равных условиях — должны стоить больше (поскольку в них может разместиться большее количество людей), а более тяжелые палатки — при прочих равных условиях — должны стоить меньше (поскольку они менее удобны при переноске и, следовательно, менее привлекательны для покупателей). В каталоге REI — компании, занимающейся продажей туристического снаряжения по почтовым заказам, — указывается цена, вес и площадь 30 видов палаток. Результаты анализа множественной регрессии для прогнозирования цены представлены в табл. 12.5.3.

   а) Стоят ли более тяжелые палатки в среднем дороже или дешевле, чем легкие, если речь идет о палатках заданного размера (т.е. площади)?

Таблица 12.5.3. Результаты множественной регрессии для цен на палатки

Уравнение регрессии имеет следующий вид:
цена = 120 + 73,2 (вес) – 7,52 (площадь).

Независимая переменная	Коэффициент	Стандартное отклонение	t-отношение	p
Константа	120,33	54,82	2,19	0,037
Вес	73,17	15,37	4,76	0,000
Площадь	–7,517	2,546	–2,95	0,006

s = 99,47 R-квадрат = 56,7% R-квадрат(коррект.) = 53,5%

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	2	349912	174956	17,68	0,000
Ошибка	27	267146	9894		
Итого	29	617058			

б) Какой показатель из приведенной компьютерной распечатки может служить ответом на п. "а"? Интерпретируйте этот показатель и укажите его единицы измерения. Является ли он значимым?

в) Соответствует ли результат из п. "а" ожиданиям относительно цены палатки, указанным при изложении условий настоящей задачи? Поясните свой ответ.

г) Стоят ли большие палатки в среднем дороже или дешевле, чем меньшие палатки, если речь идет о палатках заданного веса?

д) Какой показатель из приведенной выше компьютерной распечатки дает ответ на п. "г"? Интерпретируйте этот показатель и укажите его единицы измерения. Является ли он значимым?

е) Соответствует ли результат из п. "г" ожиданиям относительно цены палатки, указанным при изложении условий настоящей задачи? Поясните свой ответ.

8. Быстродействие компьютеров, объединенных в сеть, при возникновении перегрузок, как правило, снижается. *Время реакции* равняется интервалу с момента нажатия вами клавиши <Enter> до момента выдачи компьютером ответа на введенный вами запрос. Естественно, чем больше загрузка компьютера (в результате обращений со стороны других пользователей или выполнения какой-то другой работы), тем большим должно быть время реакции. Это время реакции (в секундах) измерялось в различные моменты времени наряду с количеством пользователей в системе и загрузкой компьютера (процент времени, в течение которого машина занята выполнением высокоприоритетных задач). Соответствующие данные представлены в табл. 12.5.4.

   а) Проанализируйте эти данные, предложив собственный комментарий по поводу взаимосвязей в трех диаграммах рассеяния, которые вы можете изобразить, рассматривая попарно указанные переменные. В частности, выглядят ли, по вашему мнению, эти взаимосвязи разумными?

   б) Вычислите корреляционную матрицу и сравните ее с взаимосвязями, которые вы наблюдаете на диаграммах рассеяния.

   в) Составьте уравнение регрессии для прогнозирования времени реакции исходя из количества пользователей и загрузки компьютера. (Для выполнения этого и последующих пунктов этой задачи вам, вероятно, придется воспользоваться компьютером.)

   г) В каких приблизительно пределах (количество секунд) для этой совокупности данных можно прогнозировать время реакции исходя из количества пользователей и загрузки компьютера?

   д) Является ли F-тест значимым? О чем это говорит вам?

   е) Являются ли значимыми коэффициенты регрессии? Интерпретируйте (в письменном виде) для каждой переменной ее поправочное влияние на время реакции.

ж) Обратите внимание, что два коэффициента регрессии очень отличаются между собой. Вычислите стандартизованные коэффициенты регрессии с целью их сравнения. Представьте в письменном виде комментарий об относительной важности количества пользователей и загрузки компьютера с точки зрения их влияния на время реакции.

9. В последние годы американская экономика более тесно интегрировалась с экономикой других стран. Но насколько тесно связанными оказываются американские и глобальные фондовые биржи в краткосрочном плане? В табл. 12.5.5 и 12.5.6 представлены соответствующие данные и результаты множественной регрессии для прогнозирования эффективности акций американских промышленных компаний по состоянию на 8 января 1993 г. на основании эффективности акций европейских промышленных компаний (X_1) и эффективности акций промышленных компаний стран Азиатско-тихоокеанского региона (X_2). Резюмируйте в письменном виде степень этой взаимосвязи. В частности, в какой мере эффективность в Европе и странах Азиатско-тихоокеанского региона объясняет эффективность в США?

10. В табл. 12.5.7 представлены некоторые результаты анализа множественной регрессии, объясняющей сумму денег, расходуемых на приобретение кухонного оборудования для приготовления пищи в домашних условиях (Y), исходя из величины дохода (X_1), уровня образования (X_2) и величины расходов на приобретение спортивного инвентаря (X_3). Все "денежные" переменные представляют общие суммы (в долларах) за прошедший год; уровень образования указан в количестве лет учебы. Рассматривается 20 наблюдений.

    а) Сколько, по вашему мнению, будет тратить человек на приобретение кухонного оборудования для приготовления пищи, если он зарабатывает $25 000 в год, проучился 14 лет и потратил в прошлом году $292 на приобретение спортивного инвентаря?

    б) Насколько удачно данное уравнение регрессии объясняет затраты на приобретение оборудования для приготовления пищи дома? В частности, на какой показатель в представленных здесь результатах следует обратить внимание и является ли он статистически значимым?

    в) С какой приблизительно точностью (в долларах за год) можно прогнозировать затраты на приобретение оборудования для приготовления пищи дома применительно к людям, охваченным настоящим исследованием?

    г) Для каждой из трех X-переменных укажите, оказывает ли она значимое влияние на затраты, связанные с приобретением оборудования для приготовления пищи дома (с учетом поправки на другие X-переменные).

Таблица 12.5.4. Время реакции компьютера, количество пользователей и уровень загрузки

Время реакции	Количество пользователей	Загрузка компьютера, %
0,31	1	20,2
0,69	8	22,7
2,27	18	41,7
0,57	4	24,6
1,28	15	20,0
0,88	8	39,0
2,11	20	33,4
4,84	22	63,9
1,60	13	35,8
5,06	26	62,3

Таблица 12.5.5. Эффективность американских, европейских и азиатско-тихоокеанских промышленных акций по состоянию на 8 января 1993 г.

	США, %	Европа, %	Азиатско-тихоокеанский регион, %
Воздушные перевозки	–0,41	0,09	–1,17
Строительные материалы	–0,29	–0,18	–1,02
Контейнеры	–0,87	–0,44	–0,74
Электротехнические компоненты	0,02	–0,29	–0,50
Заводское оборудование	–0,51	0,25	–0,59
Тяжелые конструкции	0,48	–0,10	–0,31
Тяжелое машиностроение	0,67	–0,53	–1,81
Промышленные, диверсифицированные	–0,27	–0,66	–1,21
Морской транспорт	0,16	–0,	–1,47
Контроль загрязнения окружающей среды	–0,71	0,15	–1,14
Другие промышленные услуги	–0,64	–0,49	–1,68
Железные дороги	1,10	0,09	–1,74
Транспортное оборудование	0,62	1,37	–1,16
Автомобильные перевозки	–0,12	–5,43	–0,71

Данные взяты из *The Wall Street Journal*, 1993, January 11, p. C12.

Таблица 12.5.6. Результаты множественной регрессии, касающиеся эффективности акций промышленных групп

Уравнение регрессии имеет следующий вид:

США = –0,417 + 0,029 (Европа) – 0,344 (Азиатско-тихоокеанский регион).

Независимая переменная	Коэффициент	Стандартное отклонение	t-отношение	p
Константа	–0,4166	0,4393	–0,95	0,363
Европа	0,0295	0,1138	0,26	0,801
Азиатско-тихоокеанский регион	–0,3443	0,3628	–0,95	0,363

s = 0,6119 R-квадрат = 9,0% R-квадрат(коррект.) = 0,0%

Окончание табл. 12.5.6

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	2	0,4087	0,2043	0,55	0,594
Ошибка	11	4,1189	0,3744		
Итого	13	4,5276			

Источник	DF	Seq SS
Европа	1	0,0714
Азиатско-тихоокеанский регион	1	0,3373

Таблица 12.5.7. Результаты множественной регрессии, касающиеся оборудования для приготовления пищи

Уравнение регрессии имеет следующий вид:

Y = –9,26 + 0,00137 X1 + 10,8 X2 + 0,00548 X3

Столбец	Коэффициент	Стандартное отклонение коэффициента	Т-отношение = коэф./ст.о.
	–9,26247	13,37258	–0,69264
X1	0,001373	0,000191	7,165398
X2	10,76225	0,798748	13,47389
X3	0,005484	0,025543	0,214728

S = 16,11

R-квадрат = 94,2%

11. Рассмотрим результаты множественной регрессии, представленные в табл. 12.5.8, с помощью которой предпринята попытка объяснить величину заработной платы высших руководителей 11 крупнейших топливных компаний, основываясь на объеме продаж и прибыли на собственный (акционерный) капитал (return on equity — ROE) соответствующей фирмы.[33] Например, данные по компании Exxon включают величину заработной платы, равную 1 207 (в тысячах долларов), для председателя совета директоров, величину ROE, равную 15,0 (в процентах), и объем продаж, равный 77 721 (в миллионах долларов).

    а) С какой примерно точностью (в долларах) можете вы прогнозировать величину заработной платы высших руководителей этих фирм, основываясь на их объеме продаж и ROE?

    б) Найдите прогнозируемую величину и остаточную ошибку прогнозирования для заработной платы высшего руководства Exxon, выразив оба этих показателя в долларах.

[33] Данные взяты из "Executive Compensation Scoreboard", *Business Week*, May 2, 1988, p. 76. Форма этого анализа предложена Рольфом Р. Андерсоном (Rolf R. Anderson) (личный контакт, 1989 г.).

Таблица 12.5.8. Результаты множественной регрессии для заработной платы руководителей

Уравнение регрессии имеет следующий вид:

заработная плата = 583,3609 + 0,0044 * (объем продаж) + 30,3880 * (ROE)

S = 149,3560 R^2 = 0,770379

F = 13,42005 с 2 и 8 степенями свободы

	Влияние на заработную плату	95% доверительный интервал		Проверка гипотез	Стандартная ошибка коэффициента	t-статистика
Переменная	Коэффициент	От	До	Значимый?	Стандартная ошибка	t
Константа	583,3609	389,5154	777,2064	Да	84,06136	6,939704
ROE	30,38801	11,13718	49,63883	Да	8,348146	3,640090
Объем продаж	0,004369	–0,00038	0,009128	Нет	0,002063	2,117528

    в) Если ROE интерпретируется как показатель эффективности работы фирмы, существует ли значимая связь между эффективностью и величиной заработной платы (с поправкой на объем продаж в соответствующей фирме)? Поясните свой ответ.

    г) О чем именно свидетельствует коэффициент регрессии 30,38801 для ROE?

    д) Почему объем продаж не оказывает значимого влияния на величину заработной платы (с поправкой на ROE), несмотря на достаточно большое значение 2,12 t-статистики для него?

12. Некоммерческие корпорации во многих отношениях функционируют подобно предприятиям других типов. Благотворительные организации, выполняющие больший объем операций, как правило, располагают большим штатом сотрудников, хотя у одних накладные расходы оказываются больше, чем у других. В табл. 12.5.9 представлено количество оплачиваемых штатных сотрудников благотворительных организаций, а также денежные суммы (в миллионах долларов), получаемые в результате частных пожертвований, государственных платежей и прочих источников дохода.

    а) Составьте уравнение регрессии для прогнозирования количества штатных сотрудников исходя из размеров вкладов каждого типа для этих благотворительных организаций. (Для этого вам, вероятно, придется воспользоваться компьютером.)

    б) Сколько дополнительных оплачиваемых штатных сотрудников в среднем должно, по вашему мнению, работать в благотворительной организации, которая получает в результате частных пожертвований на $5 000 000 больше, чем другая благотворительная организация (при прочих равных условиях)?

    в) С какой примерно точностью составленное вами уравнение регрессии может прогнозировать количество штатных сотрудников в этих благотворительных организациях на основании получаемых ими денежных сумм?

г) Найдите прогнозируемое количество штатных сотрудников и соответствующий остаток для American Red Cross.

д) Каким будет результат F-теста? О чем он свидетельствует?

е) Оказывают ли частные пожертвования значимое влияние на количество штатных сотрудников (при условии, что размер других пожертвований не изменяется)? Поясните свой ответ.

13. В табл. 12.5.10 представлена компьютерная распечатка части анализа, объясняющего конечную стоимость того или иного проекта на основе наиболее удачного выбора руководством фирмы величины затрат на оплату труда и сырье в момент подачи предложения о заключении контракта на выполнение этого проекта (подсчет производился на основе 25 недавно заключенных контрактов). Все переменные измеряются в долларах.

а) Какой процент вариации затрат объясняется информацией, доступной руководству фирмы в момент подачи предложения о заключении контракта?

б) С какой примерно точностью мы можем прогнозировать затраты, если нам известны другие переменные?

в) Найдите прогнозируемые затраты на выполнение проекта, оплата труда для которого планируются в размере \$9 000, а затраты на сырье — \$20 000.

г) Является ли значимым F-тест? О чем он свидетельствует?

д) Оказывает ли стоимость сырья существенное влияние на затраты?

Таблица 12.5.9. Количество штатных сотрудников и размеры взносов (в миллионах долларов) для благотворительных организаций

Благотворительная организация	Штат	Частные пожертвования	Государственные платежи	Прочие
Salvation Army	29 350	473	92	\$300
American Red Cross	22 100	341	30	602
Planned Parenthood	8 200	67	106	101
CARE	7 087	45	340	12
Easter Seals	5 600	83	51	78
Association of Retarded Citizens	5 600	28	80	32
Volunteers of America	5 000	14	69	83
American Cancer Society	4 453	271	0	37
Boys Clubs	3 650	103	9	75
American Heart Association	2 700	151	1	27
UNICEF	1 652	67	348	48
March of Dimes	1 600	106	0	6
American Lung Association	1 500	80	1	17

Данные взяты из G. Kinkead, "America's Best-Run Charities", *Fortune*, 1987, November 9, p. 146.

14. Интерпретируйте применительно к предыдущему примеру коэффициент регрессии для размера оплаты труда, оценив среднее значение конечных затрат на каждый доллар, планируемый руководством фирмы на оплату труда по соответствующему проекту.

15. Ваш коллега оказался чрезвычайно доволен: ему удалось выяснить, что значение R^2 равно 100%, указывая на то, что уравнение регрессии полностью объясняет изменчивость Y ("прибыль") на основании двух X-переменных — "доходов" и "затрат". Вы поражаете его воображение, высказав очень точный прогноз относительно величины коэффициентов регрессии.

а) Объясните, почему полученный вашим коллегой результат (R^2=100%) в данном случае является вполне ожидаемым — более того, тривиальным.

б) Каковы величины коэффициентов регрессии?

16. При вводе в эксплуатацию новой сборочной линии возникли проблемы с контролем качества. Чтобы выявить источник этих проблем, было решено воспользоваться анализом множественной регрессии. Суточный "процент брака" обозначили как переменную Y, которую требовалось прогнозировать на основании следующих переменных (которые, по предположению многих сотрудников, являлись причиной возникших проблем): "процент перегруженности" (степень перегруженности системы в сравнении со своей номинальной производительностью), "уровень буферного запаса" (степень накопления запасов между рабочими станциями) и "изменчивость на входе" (стандартное отклонение весов для важнейшего исходного компонента). На что должны быть направлены усилия руководства предприятия, если исходить из результатов множественной регрессии, представленных в табл. 12.5.11? Представьте свой ответ в форме докладной записки своему начальнику.

Таблица 12.5.10. Регрессионный анализ конечных затрат на реализацию проекта

Коэффициенты корреляции

	Затраты	Оплата работников
Оплата работников	0,684	
Стоимость сырья	0,713	0,225

Уравнение регрессии имеет следующий вид:
затраты = 13975 + 1,18 (оплата работников) + 1,64 (стоимость сырья).

Независимая переменная	Коэффициент	Стандартное отклонение	t-отношение	p
Константа	13975	4286	3,26	0,004
Оплата работников	1,1806	0,2110	5,59	0,000
Стоимость сырья	1,6398	0,2748	5,97	0,000

s = 3860 R-квадрат = 79,7% R-квадрат(коррект.) = 77,8%

Окончание табл. 12.5.10

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	2	1286267776	643133888	43,17	0,000
Ошибка	22	327775808	14898900		
Итого	24	1614043648			

Источник	DF	SEQ SS
Оплата работников	1	755914944
Стоимость сырья	1	530352896

Таблица 12.5.11. Результаты множественной регрессии для новой производственной сборочной линии

Уравнение регрессии имеет следующий вид:

брак = –1,62 + 11,7 (перегрузка) + 0,48 (буфер) + 7,29 (вход).

Независимая переменная	Коэффициент	Стандартное отклонение	t-отношение	p
Константа	–1,622	1,806	–0,90	0,381
Перегрузка	11,71	22,25	0,53	0,605
Буфер	0,479	2,305	0,21	0,838
Вход	7,290	2,287	3,19	0,005

s = 2,954 R-квадрат = 43,8% R-квадрат(коррект.) = 34,4%

Дисперсионный анализ

Источник	DF	SS	MS	F	p
Регрессия	3	122,354	40,785	4,67	0,014
Ошибка	18	157,079	8,727		
Итого	21	279,433			

17. Сменив поставщиков, вы надеетесь, что стандартное отклонение важнейшего исходного компонента удастся снизить с 0,62 до 0,38 (в среднем). Основываясь на результатах множественной регрессии из предыдущего примера, ответьте на следующий вопрос: на какое снижение уровня брака следует рассчитывать, если вы все же решите сменить поставщиков? (Уровень брака измеряется в процентных единицах; таким образом, "брак" = 5,3 соответствует 5,3% бракованных изделий.)

18. В табл. 12.5.12 представлены уровни капиталовложений и результаты, достигнутые крупнейшими производителями оптоволоконного кабеля для удаленной связи.

    а) Составьте уравнение регрессии для прогнозирования количества миль оптоволоконного кабеля исходя из величины капиталовложений.

б) Изобразите диагностическую диаграмму остатков в зависимости от прогнозируемых значений. Какую взаимосвязь вы усматриваете (если усматриваете) на этой диагностической диаграмме? Какие корректирующие действия следует предпринять в этом случае?

в) Найдите натуральный логарифм (по основанию *e*) для каждого значения данных.

г) Найдите уравнение регрессии для прогнозирования логарифма количества миль кабеля на основании логарифма капиталовложений.

д) Постройте диагностическую диаграмму для такой регрессии с использованием логарифмов. Требуются ли еще какие-то корректирующие действия? Поясните свой ответ.

е) О чем именно свидетельствует коэффициент регрессии?

ж) Найдите двусторонний 95% доверительный интервал для коэффициента регрессии (для логарифмов обеих переменных).

з) Действительно ли фирмы, осуществляющие большие капиталовложения, выпускают значительно больше миль оптоволоконного кабеля? Поясните свой ответ.

и) Для регрессии, подобной этой, при использовании логарифмов обеих переменных коэффициент регрессии, меньший 1, означает наличие экономии, обусловленной ростом масштаба производства, а коэффициент регрессии, больший 1, означает наличие дополнительных затрат на одну милю кабеля в случае более крупных проектов. Какой тип экономии предлагает

Таблица 12.5.12. Фирмы, производящие оптоволоконный кабель для удаленной связи

	Капиталовложения, млн дол.	Сетевые мили,* млн
AT&T	1 300	1 700
MCI	500	650
GTE	130	110
United Telecommunications	2 000	1 200
Fibertrak	1 200	2 400
LDX Net	110	165
Electra Communications	40	72
Microtel	60	45
Litel Telecommunications	57	85
Lightnet	500	650
SoutherNet	90	50
RCI	90	87

* *Сетевая миля* определяется как протяженность кабеля, способного передавать один речевой сигнал на расстояние в одну милю.

Данные взяты из W. B. Johnston, "The Coming Glut of Phone Lines", *Fortune*, 1985, January 7, p. 97–100. Источник: Hudson Institute.

найденный нами для этой совокупности данных коэффициент регрессии? Является ли эта экономия статистически значимой?

19. В чем, по вашему мнению, заключается суть проблемы некоторой множественной регрессии, для которой значение R^2 велико и статистически значимо, но ни для одной из X-переменных t-тест не является значимым?

20. В табл. 12.5.13 частично представлены результаты множественного регрессионного анализа, объясняющего годовые объемы продаж в 25 гастрономах на основании некоторых их характеристик. Переменная "торговая улица" равна 1, если соответствующий гастроном находится на оживленной торговой улице, и 0 — в противном случае. Переменная "посетители" равняется количеству посетителей гастронома за год.

    а) С какой примерно точностью (в долларах) можно прогнозировать объем продаж на основе данной регрессионной модели?

    б) Найдите прогнозируемый объем продаж для гастронома, находящегося на оживленной торговой улице и имеющего 100 000 покупателей за год.

    в) Оказывают ли эти независимые переменные существенное влияние на объем продаж? Поясните свой ответ.

    г) О чем именно свидетельствует коэффициент регрессии для количества покупателей?

    д) Оказывает ли место расположения гастронома (оживленная торговая улица или более тихое место) существенное влияние на объем продаж, если сравнивать два гастронома с одинаковым количеством покупателей за год? Поясните кратко, почему такое влияние может иметь место.

    е) Какой (примерно) дополнительный годовой объем продаж обеспечивает себе гастроном, находящийся на оживленной торговой улице, в сравнении с подобным ему гастрономом, расположенным в более тихом месте?

21. Ценообразование, как правило, — непростая задача. Заниженная цена обычно способствует повышению объема продаж, однако прибыль в расчете на одну продажу в этом случае оказывается ниже. Завышенная цена обеспечивает более высокую прибыль в расчете на одну продажу, однако в целом объем продаж снижается. Обычно фирма стремится выбрать такую цену, которая максимизирует общую прибыль, однако при этом следует учитывать существование значительной неопределенности в отношении спроса. В табл. 12.5.14 представлены гипотетические результаты исследования прибыли на сопоставимых тестовых рынках одинакового размера, где меняется лишь цена.

    а) Составьте уравнение регрессии в следующей форме: прогнозируемая прибыль = $a + b$(цена).

    б) Проверьте, значима ли данная регрессия. Можно ли считать логически обоснованным полученный вами результат?

    в) С какой примерно точностью (в долларах) можно прогнозировать прибыль на основании цены, если воспользоваться предложенным здесь способом?

Таблица 12.5.13. Результаты множественной регрессии для годового объема продаж в гастрономах

Уравнение регрессии имеет следующий вид:

объем продаж = –36589 + 209475 (торговая улица) + 10,3 (покупатели).

Независимая переменная	Коэффициент	Стандартное отклонение	t-отношение	p
Константа	–36589	82957	–0,44	0,663
Торговая улица	209475	77040	2,72	0,013
Покупатели	10,327	4,488	2,30	0,031

s = 183591 R-квадрат = 39,5% R-квадрат(коррект.) = 34,0%

Таблица 12.5.14. Цена и прибыль на тестовых рынках (в долларах)

Цена	Прибыль
8	6 486
9	10 928
10	15 805
11	13 679
12	12 758
13	9 050
14	5 702
15	–109

    г) Проанализируйте диагностическую диаграмму и выясните, присутствует ли в ней еще какая-нибудь структура, которая помогла бы объяснить прибыль на основании цены. Опишите структуру, которую вам удалось выявить.

    д) Создайте еще одну X-переменную, используя квадрат цены, и составьте уравнение множественной регрессии для прогнозирования прибыли исходя из цены и ее квадрата.

    е) С какой примерно точностью (в долларах) можно прогнозировать прибыль на основании цены, если воспользоваться двумя указанными выше X-переменными?

    ж) Проверьте, объясняют ли взятые вместе цена и ее квадрат значимую долю вариации прибыли.

    з) Найдите цену, при которой прогнозируемая прибыль достигает максимума. Сравните полученное значение с ценой, при которой наблюдаемая прибыль достигла наивысшего значения.

22. В табл. 12.5.15 представлены результаты анализа множественной регрессии, целью которой является объяснение уровня заработной платы высших должностных лиц на основании объемов продаж в их фирме и на основании

Таблица 12.5.15. Результаты множественной регрессии, касающиеся заработной платы руководителей фирм

Уравнение регрессии имеет следующий вид:

заработная плата = 931,8383
+0,01493 * (объем продаж)
–215,747 * (аэрокосмическая отрасль)
–135,550 * (банковская сфера)
–303,774 * (полезные ископаемые)

$S = 401,8215$

$R^2 = 0,423469$

Переменная	Коэффициент	Стандартная ошибка
Константа	931,8383	163,8354
Объем продаж	0,014930	0,003047
Аэрокосмическая отрасль	–215,747	222,3225
Банковская сфера	–135,550	177,0797
Полезные ископаемые	–303,774	187,4697

промышленной группы.[34] Переменная Y представляет величину заработной платы руководителя фирмы (в тысячах долларов). Переменная X_1 представляет объем продаж в соответствующей фирме (в миллионах долларов). X_2, X_3 и X_4 являются индикаторными переменными, которые представляют соответственно промышленные группы аэрокосмической отрасли, банковской сферы и отрасли добычи полезных ископаемых (группа добычи полезных ископаемых включает крупные нефтяные компании). Индикаторная переменная для базовой промышленной группы — автомобилестроение — не включена. Совокупности данных содержат $n = 49$ наблюдений.

а) Оказывают ли объем продаж и промышленная группа значимое влияние на размер заработной платы высших руководителей фирм?

б) Каково оцениваемое влияние каждого дополнительного миллиона долларов объема продаж на уровень заработной платы руководителя фирмы (с поправкой на промышленную группу)?

в) Является ли статистически значимой оцененная вами разница в уровнях заработной платы руководителей, вызванная разницей в объемах продаж (см. п. "б")? Какие практические выводы можно сделать, исходя из этой разницы в уровнях заработной платы?

г) Учитывая соответствующий коэффициент регрессии, ответьте, насколько больше (или, наоборот, меньше) оплачивается труд руководителя банка в сравнении с оплатой труда руководителя автомобилестроительной компании сопоставимого масштаба?

[34] Использованные данные взяты из статьи "Executive Compensation Scoreboard", *Business Week*, 1988, May 2, p. 57.

д) Является ли статистически значимой оцененная вами в п. "г" разница в уровнях заработной платы руководителей фирм банковской и автомобилестроительной отраслей? Какие практические выводы можно сделать исходя из этой разницы в уровнях заработной платы?

23. Рассмотрите пример с затратами на размещение рекламных объявлений в журналах из раздела 12.1.

а) Какая из X-переменных наименее полезна с точки зрения объяснения величины тарифа на размещение рекламы в журналах? Поясните свой ответ.

б) Выполните еще раз регрессионный анализ, отбросив эту X-переменную.

в) Сравните следующие результаты без использования X-переменной с результатами в случае использования X-переменной: F-тест, R^2, коэффициенты регрессии и t-статистики.

24. Рассмотрите ставки процента по ценным бумагам с различными сроками погашения (соответствующие данные представлены в табл. 12.5.16).

а) Найдите уравнение регрессии для прогнозирования долгосрочной ставки процента (долгосрочные казначейские обязательства) на основании двух других ставок процента (с меньшими сроками погашения).

Таблица 12.5.16. Ставки процента

Год	Федеральные фонды (однодневная процентная ставка)	Казначейские векселя (трехмесячная процентная ставка)	Долгосрочные казначейские обязательства (десятилетняя процентная ставка)
1980	13,35	11,39	11,43
1981	16,39	14,04	13,92
1982	12,24	10,60	13,01
1983	9,09	8,62	11,1
1984	10,23	9,54	12,46
1985	8,10	7,47	10,62
1986	6,80	5,97	7,67
1987	6,66	5,78	8,39
1988	7,57	6,67	8,85
1989	9,21	8,11	8,49
1990	8,10	7,50	8,55
1991	5,69	5,38	7,86
1992	3,52	3,43	7,01
1993	3,02	3,00	5,87
1994	4,21	4,25	7,69
1995	5,83	5,49	6,57
1996	5,30	5,01	6,44

Данные взяты из таблиц 806 и 807 Бюро переписи населения США, *Statistical Abstract of the United States: 1997* (117th edition.) Washington, DC, 1997.

б) Создайте новую переменную ("взаимодействие"), перемножив два вида ставок процента с меньшими сроками погашения. Найдите уравнение регрессии для прогнозирования долгосрочной ставки процента (долгосрочные казначейские обязательства) на основании двух других ставок процента (с меньшими сроками погашения) и переменной "взаимодействие".

в) Проверьте, есть ли какое-либо взаимодействие двух видов ставок процента с меньшими сроками погашения, которое являлось бы частью взаимосвязи между краткосрочными и долгосрочными ставками процента.

Упражнения с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении А.

1. Рассмотрите прогнозирование годовой заработной платы исходя из возраста и стажа работы.

   а) Найдите и интерпретируйте уравнение регрессии и коэффициенты регрессии.

   б) Найдите и интерпретируйте стандартную ошибку оценки.

   в) Найдите и интерпретируйте коэффициент детерминации.

   г) Является ли данная модель значимой? О чем это свидетельствует?

   д) Проверьте каждый коэффициент регрессии на значимость и интерпретируйте полученные результаты.

   е) Найдите и интерпретируйте стандартизованные коэффициенты регрессии.

   ж) Проанализируйте диагностическую диаграмму и выявите серьезные проблемы — если таковые действительно имеются.

2. Продолжайте использовать прогнозы годовой заработной платы служащих исходя из их возраста и стажа работы.

   а) Найдите прогнозируемую годовую заработную плату и ошибку прогнозирования для служащего под номером 33 и сравните полученный результат с его фактической годовой заработной платой.

   б) Найдите прогнозируемую годовую заработную плату и ошибку прогнозирования для служащего с номером 52 и сравните полученный результат с его фактической годовой заработной платой.

   в) Найдите прогнозируемую годовую заработную плату и ошибку прогнозирования для самого высокооплачиваемого служащего и сравните полученный результат с его фактической годовой заработной платой. О чем свидетельствуют результаты этого сравнения?

   г) Найдите прогнозируемую годовую заработную плату и ошибку прогнозирования для самого низкооплачиваемого служащего и сравните полученный результат с его фактической годовой заработной платой. О чем свидетельствуют результаты этого сравнения?

3. Рассмотрите прогнозирование годовой заработной платы исходя из одного лишь возраста (сравните с упражнением 1, где в качестве X-переменной также использовался стаж работы).

   а) Найдите уравнение регрессии для прогнозирования годовой заработной платы исходя из возраста служащего.

   б) Используя результаты, полученные в п. "а" упражнения 1 и настоящего упражнения, сравните влияние возраста на годовую заработную плату с поправкой на стаж работы и без такой поправки.

   в) Проверьте, оказывает ли возраст служащего значимое влияние на годовую заработную плату с поправкой на рабочий стаж и без такой поправки. Кратко обсудите полученные результаты.

4. Теперь проанализируйте влияние пола служащего на его годовую заработную плату с поправками на его возраст и стаж работы и без таких поправок.

   а) Найдите среднюю годовую заработную плату для мужчин и для женщин и сравните полученные значения.

   б) Используя двусторонний тест на уровне 5%, выясните, зарабатывают ли мужчины значимо больше, чем женщины. (Возможно, чтобы найти подходящий для данного случая тест, вам придется вернуться к материалу главы 10.)

   в) Найдите уравнение множественной регрессии для прогнозирования годовой заработной платы исходя из возраста, стажа работы и пола служащего, воспользовавшись индикаторной переменной для пола (выбрав для женщин значение переменной, равное 1).

   г) Проанализируйте и интерпретируйте коэффициент регрессии для пола служащего.

   д) Оказывает ли пол служащего значимое влияние на годовую заработную плату — с поправкой на возраст и стаж работы?

   е) Сравните и обсудите результаты, полученные вами в пп. "б" и "д" настоящего примера.

5. Теперь проанализируйте влияние уровня подготовки служащего на его годовую заработную плату с поправками на его возраст и стаж работы и без таких поправок.

   а) Найдите среднюю годовую заработную плату для каждого из трех уровней подготовки и сравните полученные результаты.

   б) Найдите уравнение множественной регрессии для прогнозирования годовой заработной платы исходя из возраста, стажа работы и уровня подготовки служащего, воспользовавшись индикаторными переменными для уровня подготовки. Используйте уровень А в качестве базы и не используйте соответствующую индикаторную переменную.

   в) Проанализируйте и интерпретируйте коэффициент регрессии для каждой индикаторной переменной, которая соответствует одному из уровней подготовки.

   г) Оказывает ли уровень подготовки служащего значимое влияние на его годовую заработную плату — с поправкой на возраст служащего и его стаж работы?

д) Сравните и обсудите среднюю разницу в заработной плате между уровнями подготовки A и C — как с поправкой на возраст и стаж работы, так и без нее.

6. Рассмотрите прогнозирование годовой заработной платы исходя из возраста, стажа работы и терма взаимодействия.

   а) Создайте новую переменную ("взаимодействие"), умножив для каждого служащего его возраст на стаж работы.

   б) Найдите уравнение регрессии для прогнозирования годовой заработной платы исходя из возраста служащего, его стажа работы и терма взаимодействия.

   в) Проверьте, насколько значимым является данное взаимодействие, воспользовавшись t-тестом для коэффициента регрессии, относящегося к переменной взаимодействия.

   г) Какое среднее влияние на годовую заработную плату оказывает каждый дополнительный год стажа работы у 40-летнего служащего?

   д) Какое среднее влияние на годовую заработную плату оказывает каждый дополнительный год стажа работы у 50-летнего служащего?

   е) Интерпретируйте взаимодействие между возрастом и рабочим стажем, сравнив свои ответы на пп. "г" и "д" настоящего упражнения.

Проект



Найдите в Internet, в вашей библиотеке, в газете или в журнале многомерную совокупность данных, касающуюся вашей работы или интересов вашего бизнеса. Размер выборки должен составлять $n = 25$ или больше; кроме того, F-тест, а также по меньшей мере один из t-тестов для этой совокупности данных должны быть значимыми.

а) Выберите зависимую переменную (Y) и кратко поясните причины, которые заставили вас остановить свой выбор именно на ней.

б) Исследуйте и прокомментируйте диаграммы рассеяния, представляющие собой зависимость переменной Y от каждой из X- переменных.

в) Вычислите и кратко интерпретируйте матрицу корреляций.

г) Составьте уравнение регрессии.

д) Для двух элементарных единиц вашей совокупности данных вычислите прогнозируемые значения Y и остатки.

е) Интерпретируйте каждый коэффициент регрессии и его доверительный интервал.

ж) Какие из коэффициентов регрессии являются значимыми? Какие из них не являются значимыми (если таковые имеются)? Имеют ли смысл полученные вами результаты?

з) Укажите, что нового относительно влияния X-переменных на переменную Y вы узнали из анализа множественной регрессии.

Ситуация для анализа

Контроль качества продукции

По поводу того, почему так много изделий после их изготовления приходится переделывать или вообще выбрасывать, единого мнения не существует. Некоторые утверждают, что все дело в температуре соответствующего производственного процесса, которую необходимо поддерживать постоянной (ее колебания возможны лишь в допустимых пределах). Другие заявляют, что главное — плотность материала, из которого производятся изделия: если применять более прочный материал, проблемы исчезнут сами собой. А есть еще и Оле, который в свое время предупреждал, что нельзя использовать производственное оборудование в режимах, не предусмотренных его техническими характеристиками. Это условие можно выполнить довольно просто: нужно просто снизить производительность системы; однако в этом случае придется смириться с ростом себестоимости продукции. Интересно отметить, что многие работники утренней смены полагают, что проблема заключается в "низкой квалификации работников дневной смены" (впрочем, то же самое работники дневной смены говорят и о своих коллегах).

После того как производственный процесс был автоматизирован (компьютеры, управляющие процессом, объединены в сеть, а на каждой рабочей станции установлены устройства для считывания штрих-кодов), удалось наладить сбор необходимых данных. Решением проблемы качества изделий поручено заняться вам. После того как ваш помощник организовал данные в виде четырехчасовых блоков, а затем ввел переменную, обозначающую рабочую смену до или после обеда, вы нашли на своем рабочем столе следующую записку, к которой прилагалась распечатка данных, уже загруженных в компьютерную сеть.

Интересные новости! Используются следующие переменные.

- *Переменная "температура" фактически содержит результат измерения изменчивости температуры в виде стандартного отклонения в течение времени измерения.*
- *Переменная "плотность" содержит плотность материала конечного продукта.*
- *Переменная "производительность" содержит значения производительности данного процесса.*
- *AM/PM представляет собой индикаторную переменную (равняется 1 в утреннюю смену и 0 — в дневную).*
- *Переменная "дефект" содержит среднее количество дефектов на каждые 1 000 произведенных изделий.*

Температура	Плотность	Производительность	AM/PM	Дефект
0,97	32,08	177,7	0	0,2
2,85	21,14	254,1	0	47,9
2,95	20,65	272,6	0	50,9
2,84	22,53	273,4	1	49,7
1,84	27,43	210,8	1	11,0

Окончание таблицы

Температура	Плотность	Производительность	AM/PM	Дефект
2,05	25,42	236,1	1	15,6
1,50	27,89	219,1	0	5,5
2,48	23,34	238,9	0	37,4
2,23	23,97	251,9	0	27,8
3,02	19,45	281,9	1	58,7
2,69	23,17	254,5	1	34,5
2,63	22,70	265,7	1	45,0
1,58	27,49	213,3	0	6,6
2,48	24,07	252,2	0	31,5
2,25	24,38	238,1	0	23,4
2,76	21,58	244,7	1	42,2
2,36	26,30	222,1	10	13,4
1,09	32,19	181,4	1	0,0
2,15	25,73	241,0	0	20,6
2,12	25,18	226,0	0	15,9
2,27	23,74	256,0	0	44,4
2,73	24,85	251,9	1	37,6
1,46	30,01	192,8	1	2,2
1,55	29,42	223,9	1	1,5
2,92	22,50	260,0	0	55,4
2,44	23,47	236,0	0	36,7
1,87	26,51	237,3	0	24,5
1,45	30,70	221,0	1	2,8
2,82	22,30	253,2	1	60,8
1,74	28,47	207,9	1	10,5

Вы, естественно, решаете выполнить множественную регрессию для прогнозирования частоты появления дефектов на основании всех независимых (объясняющих) переменных. Идея такого подхода заключается в том, чтобы понять, какие переменные связаны с появлением дефектов (если такие переменные действительно имеются). Кроме того, вы полагаете: если какая-то переменная помогает прогнозировать появление дефектов, значит, существует возможность контролировать (снижать) частоту появления дефектов, изменяя эту переменную. Ниже приведены результаты регрессии, вычисленные с помощью электронных таблиц.[35]

[35] Обратите внимание, что число можно представить в так называемой научной нотации; таким образом, 2,36E–5 означает $(2,36)(10^{-5}) = 0,0000236$. Указанный способ записи чисел можно представить себе следующим образом: E–5 означает, что десятичную точку нужно сдвинуть на 5 позиций влево.

Итоговая распечатка

Статистические показатели регрессии

Множественный R	0,948
R-квадрат	0,899
R-квадрат корректированный (с поправкой)	0,883
Стандартная ошибка	6,644
Наблюдений	30

ANOVA

	df	SS	MS	F	p-значение
Регрессия	4	9825,76	2456,44	55,65	4,37E–12
Остаток	25	1103,54	44,14		
Итого	29	10929,29			

	Коэффициент	Стандартная ошибка	t	p	95% нижняя граница	95% верхняя граница
Сдвиг	–28,756	64,170	–0,448	0,658	–160,915	103,404
Температура	26,242	9,051	2,899	0,008	7,600	44,884
Плотность	–0,508	1,525	–0,333	0,742	–3,649	2,633
Производительность	0,052	0,126	0,415	0,682	–0,207	0,311
AM/PM	–1,746	0,803	–2,176	0,039	–3,399	–0,093

На первый взгляд, выводы из этих результатов достаточно очевидны. Но так ли это на самом деле?

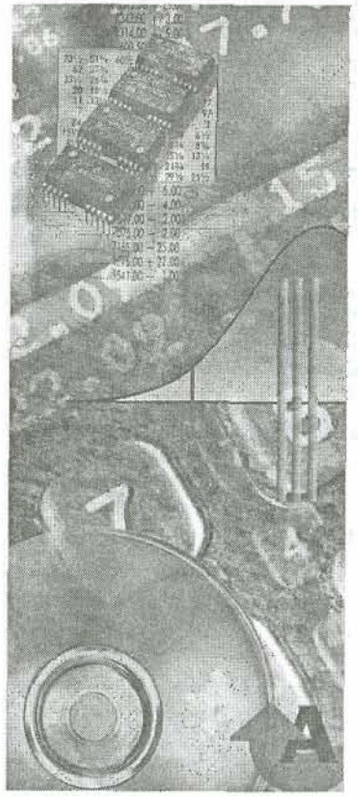
Вопросы для обсуждения

1. В чем заключаются "очевидные выводы" из проверок гипотез в распечатке результатов регрессии?
2. Проанализируйте приведенные данные. Не находите ли вы в них чего-то такого, что заставляет усомниться в результатах регрессии? Если потребуется, выполните дальнейший анализ.
3. Что бы вы порекомендовали предпринять? Почему?

ГЛАВА 13

Составление отчетов: представление результатов множественной регрессии

Умение грамотно изложить (представить) результаты проделанной работы — важная составляющая профессиональной деятельности в большинстве областей. Менеджер использует соответствующие коммуникационные стратегии для мотивации тех, кто представляет ему на рассмотрение результаты своей работы (т.е. своих подчиненных), чтобы убедить своего начальника в важности полученных результатов, убедить в чем-то своих потенциальных клиентов, воздействовать нужным образом на поставщиков своей фирмы и т.п.



Статистические отчеты нередко помогают наиболее объективным и удобным способом довести до сведения других людей важнейшую информацию о соответствующей ситуации.[1] Они позволяют вам довести свою точку зрения до самой широкой аудитории. Доверие к вам со стороны других людей в этом случае может повыситься, поскольку им должно быть совершенно очевидно, что для того, чтобы представить публике "полную картину" ситуации, вам пришлось приложить определенные усилия, тщательно проанализировав все имеющиеся у вас сведения. Ниже приведено несколько примеров отчетов, которые включают статистическую информацию.

Первый. *Обследование рынка.* Руководство вашей фирмы решает, запускать ли в производство новый продукт. Обследование рынка позволяет получить важную базовую информацию о потенциальных потребителях продукции: их симпатиях и антипатиях, суммах, которые они готовы выложить за понравившийся им товар, поддержке со стороны производителя, на которую они рассчитывают, и т.д. Чтобы вам легче было уяснить особенности своих потенциальных потребителей, в соответствующий отчет можно включить анализ множественной регрессии, который — *на основании таких* характеристик, как доход и промышленная группа, — помог бы вам понять, действительно ли эти потребители готовы приобрести выпускаемый вами продукт. Цель отчета — обеспечить базовую информацию для принятия решения о выпуске продукции. Такой отчет предназначен для руководителей среднего и высшего звена, в том числе и тех, кто будет принимать окончательное решение.

Второй. *Рекомендации по усовершенствованию производственного процесса.* Учитывая *возможное наличие* местных и зарубежных конкурентов и при условии, что срок действия вашего патента истекает в следующем году, вы все же хотите остаться на рынке в качестве производителя с низкой себестоимостью выпускаемой продукции. Эта цель вполне достижима, поскольку, в конце концов, вы располагаете намного большим опытом, чем ваши конкуренты. Соответствующий отчет на основе данных, собранных в ходе реального производства, а также в результате экспериментов с альтернативными процессами, включает информацию о применении различных подходов и показатели ожидаемой экономии при различных сценариях развития событий. В этом отчете могут содержаться результаты множественной регрессии, позволяющие выявить важные факторы и предложить способы их корректировки. Цель такого отчета — помочь снизить расходы. Такой отчет *предназначен для руководителей* среднего и высшего звена, принимающих решение о том, каким предложениям отдать предпочтение.

Третий. *Анализ практики приема на работу новых сотрудников и политики в области заработной платы.* Руководство вашей фирмы анализирует — в форме периодического обзора или в ответ на обвинения в несправедливом подходе к отдельным работникам — свою текущую политику в области управления кадрами. Анализ множественной регрессии можно использовать для объяснения величины заработной платы исходя из возраста, стажа работы, пола, квалификации и других характеристик работника. Сравнивая официальную политику фирмы с результатами регрессионного анализа, можно выяснить, насколько эффективно фирма реализует свои цели. Проверяя коэффициент регрессии для пола сотрудника, *можно выяснить, наблюдаются* ли в вашей фирме какие-то проявления дискриминации работников по признаку пола. Среди целей подобного исследования может быть оказание помощи фирме в достижении ею целей своей кадровой политики и, возможно, защита ее руководства от обвинений в дискриминации работников. Подобный отчет может быть *предназначен* для руководителей среднего и высшего звена, пытающихся внести определенные коррективы в кадровую политику фирмы; кроме того, он может служить в качестве одного из аргументов в ходе судебного разбирательства.

Четвертый. *Определение структуры затрат.* Пытаясь управлять затратами, полезно знать, каковы они на самом деле. В частности, необходимо знать, как колеблется спрос и производство, какой компонент ваших суммарных затрат

[1] Возможны, разумеется, и другие применения статистики. Обратитесь, например, к книге Huff D. *How to Lie with Statistics* (New York: Norton, 1954) ("Как обманывать с помощью статистики").

можно рассматривать как фиксированные затраты и что в таком случае представляют собой переменные затраты на каждое изделие, производимое вашей фирмой. Анализ множественной регрессии предоставит в ваше распоряжение оценки вашей структуры затрат, основанные на реальном производственном опыте. Цель данного исследования — уяснение структуры затрат и управление ею. Реальными "потребителями" подобного отчета являются менеджеры, отвечающие за калькуляцию затрат и управление ими.

Пятый. *Тестирование продукции.* Ваша фирма, как вам представляется, выпускает наилучшую продукцию в соответствующей категории. Одним из способов убедить в этом других является представление статистических результатов на основании объективного тестирования продукции. Фирмы, выпускающие зубную пасту, в настоящее время широко пользуются этим приемом ("...зарекомендовала себя эффективным средством против кариеса, поддерживающим зубы в идеальном порядке..."). Для этого могут использоваться различные методы проверки статистических гипотез — от t-теста для независимых выборок до множественной регрессии. Цель подобного исследования — доказать превосходство вашей продукции. Его результаты предназначены для ваших потенциальных клиентов. Форма такого отчета может быть самой разной: от подходящей цитаты на упаковке до абзаца в информационной брошюре или даже 200-страничного отчета, который подается в соответствующее государственное агентство и предоставляется по требованию.

Допустим, что основной целью использования в отчете полученных вами статистических результатов является их *изложение* для других людей. Проявите внимание к своим читателям и объясните им, что нового вы узнали в результате проведенного исследования; используйте при этом язык, понятный вашим читателям. Ваша работа наверняка произведет на них большее впечатление, если они действительно поймут ее, а не потеряются в дебрях технической терминологии и подробностях, понятных лишь узкому кругу специалистов.

Отчеты составляются по разным причинам и для различных потребителей. Уяснив, кто именно будет потребителем вашего отчета, вы существенно облегчите свою задачу (составление отчета), поскольку сможете представить себе реальную аудиторию, к которой вы обращаетесь с вполне реальной целью. Уяснение конкретной цели поможет вам сузить тематику своего отчета и позволит сконцентрироваться на вопросах, имеющих непосредственное отношение к этой цели. Определение конкретной аудитории поможет вам подобрать соответствующий стиль написания отчета и уровень его детализации.

13.1. Как организовать свой отчет

Способ организации отчета зависит от вашей цели и от аудитории, на которую вы ориентируетесь. В этом разделе приведен обзор основных частей статистического отчета, которые вы можете модифицировать, исходя из своей конкретной цели. Предлагаемая нами форма типового отчета состоит из шести частей.

1. *Реферат* представляет собой абзац в самом начале отчета, где описываются наиболее важные факты и выводы из проделанной вами работы.

2. *Введение* состоит из нескольких абзацев, в которых вы описываете "предысторию вопроса", цель исследования и данные, с которыми вы работали.
3. *Раздел "Анализ и методы"* позволяет вам интерпретировать исходные данные, включив в отчет их графическое представление, итоговые статистические показатели и результаты, которые вы объясняете по ходу дела.
4. Цель *раздела "Выводы и резюме"* заключается в том, чтобы представить общую картину ситуации, собрав воедино все важнейшие мысли, которые вы хотели бы довести до сведения своей аудитории.
5. *Раздел "Ссылки"* содержит указания на сведения, полученные вами из внешних источников. Назначение этого раздела — предоставить возможность своим читателям в случае необходимости самостоятельно обратиться к этим источникам. Эти ссылки можно указывать по ходу текста (на соответствующих страницах) или собрать их в самостоятельный раздел.
6. *Приложение* должно содержать весь вспомогательный материал, который, на ваш взгляд, является достаточно важным, чтобы присутствовать в отчете, и, тем не менее, не столь уж значительным, чтобы включать его в основной текст отчета.

Кроме того, ваш отчет может содержать *титульную страницу* и *оглавление*. **Титульная страница** идет первой и включает название отчета, фамилию и должность лица, для которого вы подготовили этот отчет, вашу фамилию и должность (как основного исполнителя) и дату. **Оглавление** следует после реферата и отражает структуру отчета (с указанием номеров страниц).

Хорошо организованный отчет характеризуется ясностью и простотой. Помните: ваша цель — донести нужную информацию до соответствующей аудитории, чтобы люди поняли суть проделанной вами работы. Не пытайтесь заинтриговать своих читателей, оставляя самое интересное и важное на последние страницы отчета. Напротив, самые важные результаты необходимо сообщить уже в начале отчета, а затем на последующих страницах лишь раскрывать детали исследования. Ваши читатели оценят такой подход по достоинству, поскольку они не меньше вашего дорожат собственным рабочим временем. К тому же такой подход поможет вам донести нужную информацию до тех читателей, которые в силу своей занятости смогут ознакомиться лишь с частью вашего отчета.

Составляя свой отчет, пользуйтесь планом, в котором каждый абзац будущего отчета отражается не более чем одной строкой. Такой план — превосходный способ держать в голове общую картину ситуации в то время, когда приходится оттачивать отдельные формулировки текста отчета.

Абзац, содержащий реферат

Реферат представляет собой абзац в самом начале отчета, содержащий описание наиболее важных фактов и выводов из проделанной вами работы (менее существенные детали здесь опускаются). Пользуясь при написании этого абзаца простым, понятным ("нетехническим") языком, вы ориентируете своего читателя на суть рассматриваемой проблемы и поясняете свой вклад в ее понимание и решение. Вы, в принципе, пытаетесь "втиснуть" весь свой отчет в единственный абзац.

Некоторые люди, особенно те, кто считают себя "технически ориентированными", могут быть недовольны тем, что сотни страниц проведенных ими аналитических исследований и без того уже сокращены до 15-страничного отчета и что совершенно невозможно (да и несправедливо) пытаться "ужать" столь ценную работу в один абзац. Однако следует учитывать, что с вашим отчетом, возможно, будут знакомиться люди, чье рабочее время ценится очень высоко. Эти люди, скорее всего, прочитают лишь реферат отчета, и если вы хотите довести нужную информацию до их сведения, то лишь облегчите свою (и их) задачу тем, что постараетесь изложить суть своего исследования в одном кратком абзаце.

Несмотря на то что реферат идет первым, нередко бывает гораздо проще написать его в самую последнюю очередь — после того, как будут составлены все остальные части вашего отчета. Только в этот момент вы можете быть полностью уверены в том, что именно необходимо резюмировать!

Вводная часть

Введение состоит из нескольких абзацев, в которых вы описываете "предысторию вопроса", цель исследования и данные, с которыми вы работали. Постарайтесь изложить материал вводной части простым, "нетехническим" языком — так, будто вы хотите ознакомить с ним умного человека, не очень-то знакомого с подробностями исследуемой ситуации. Ознакомившись с рефератом и введением, ваш читатель должен полностью ориентироваться в исследуемой ситуации. Все остальное — лишь детали.

Во введении можно повторить материал из реферата. Можно даже непосредственно "скопировать" несколько предложений из реферата, использовав их в качестве своеобразного "введения к введению".

Раздел "Анализ и методы"

В разделе про **анализ и методы** вы интерпретируете исходные данные, включая в отчет их графическое представление, итоговые статистические показатели и результаты, которые вы объясняете по ходу дела. Здесь у вас появляется возможность изложить некоторые подробности, о которых в реферате и введении вы упоминали лишь вскользь.

Постарайтесь выбрать главное, отделив его от второстепенного. Вам, вероятно, придется опустить немалую часть своих аналитических выкладок. Внимательный аналитик, как правило, исследует все возможности — так, "на всякий случай", — проверяя различные предположения и правильность основного подхода. Но многие из этих результатов предназначены для "отдельной папки" — архива всех наблюдений. Из этой папки следует отобрать лишь то, что существенно для ситуации, рассматриваемой в вашем отчете. Если, например, для некоторой группы диаграмм рассеяния характерна обычная линейная структура, в отчет можно включить лишь одну из них, сопроводив ее комментарием о том, что другие оказались практически такими же. Отбирайте материалы, непосредственно относящиеся к поставленной вами цели.

Чтобы наилучшим образом донести до читателя свою точку зрения, графический материал желательно помещать на страницу с текстом, в котором этот материал обсуждается. В этом вам поможет компьютер. Альтернативным вариантом является использование копировального аппарата с уменьшением копий, который позволяет вставлять небольшие графические изображения непосредственно на страницу с текстом.

Ниже приведен перечень вопросов, которые желательно осветить в разделе "Анализ и методы"; эти вопросы упорядочены в соответствии с четырьмя основными задачами статистики.

1. *Планирование исследования.* Если есть какие-то важные аспекты, касающиеся способов получения вами данных, которые невозможно было осветить во введении, их можно включить либо в этот раздел, либо в приложение.
2. *Исследование данных.* Сообщите своему читателю, *что* вы обнаружили в ходе исследования данных. Сюда же можно включить некоторые графические материалы (гистограммы, блочные диаграммы или диаграммы рассеяния), если они помогают читателю понять суть излагаемой вами темы. Чтобы объяснить причины использования того или иного преобразования, в отчет можно включить какую-нибудь чрезвычайно асимметричную гистограмму или диагностическую диаграмму со структурой — это поможет объяснить читателю ход ваших рассуждений. Если по ходу исследований у вас возникли проблемы с выбросами (резко отклоняющимися значениями), здесь вы можете упомянуть об этом и показать, как вам удалось справиться с этой проблемой.
3. *Оценка.* Представьте соответствующие случаю статистические показатели и укажите, о чем они свидетельствуют с точки зрения исследуемой вами ситуации в бизнесе. Это могут быть средние значения (указывающие типичное значение переменной), стандартные отклонения (возможно, указывающие риск), корреляции (указывающие силу взаимосвязи) или коэффициенты регрессии (указывающие влияние одного фактора на другой — с определенными поправками). Возможно, понадобится также включить в эти оценки меры вариации (изменчивости), чтобы ваши читатели могли оценить качество предоставленной вами информации. В их число можно включить — когда это возможно и уместно — стандартные ошибки и доверительные интервалы оценок, а также коэффициент детерминации R^2 и стандартную ошибку оценки для регрессионного анализа.
4. *Проверки гипотез.* В случае необходимости сообщите своему читателю, "имеют ли в действительности место" те влияния, которые вам удалось оценить, сравнивая их, например, с заданным эталонным значением 0. Получив статистически значимый результат, вы имеете право объяснить его. Если же какая-то оценка *не* является статистически значимой, вы *не* должны заниматься ее интерпретацией.[2] Выполняя проверку, вы показываете своему читателю, что ваши утверждения имеют веские основания.

[2] Если, например, оказалось, что какой-то из коэффициентов регрессии равен −167,35, но при этом не отличается значимо от 0, то на самом деле вы не можете быть уверены даже в том, что этот коэффициент регрессии — отрицательное число. Поскольку истинный эффект (в генеральной совокупности) может быть положительным, а не отрицательным, не пытайтесь "объяснить", почему он отрицателен, — вы можете заблуждаться!

Раздел "Выводы и резюме"

К этому моменту ваш читатель уже в какой-то степени знаком с подробностями вашего проекта. Цель раздела **Выводы и резюме** заключается в том, чтобы вернуть читателя к общей картине исследуемой ситуации, собрав воедино все важнейшие мысли, которые вы хотели бы довести до сведения своей аудитории. В то время как цель реферата в первую очередь заключается в "начальной ориентации" потребителей вашего отчета, раздел "Выводы и резюме" может отразить те подробности, которые вы представили в предыдущих разделах. Помните: в то время как часть читателей ознакомится со всем материалом, предшествующим этому разделу, другая их часть может ограничить знакомство с вашим отчетом чтением только этого раздела.

Постарайтесь, в частности, сообщить своим читателям, что вам удалось установить в результате проведенного анализа. Что вообще заставило вас выполнить его? В чем ценность полученных вами результатов? Как вы ответили на поставленные вопросы?

Включение ссылок

Если вы используете в своем отчете какие-то тексты, данные или идеи из сторонних источников, необходимо делать ссылки на эти источники. Ссылка указывает на тип материала, заимствованного вами из стороннего источника информации, и позволяет читателю самостоятельно обратиться — при необходимости — к этому источнику. Каждую такую ссылку можно оформить в виде сноски (подстрочного примечания) на странице текста, где эта ссылка указана, или же собрать все ссылки в специальный раздел.

Ссылка должна содержать достаточно сведений для того, чтобы заинтересовавшийся ею читатель действительно мог найти информацию, которой вы воспользовались. Недостаточно просто сообщить фамилию автора или указать "Министерство торговли США". Полная ссылка должна включать и такую информацию, как дата, том, страница и название издательства. Даже такое предложение, как "объем продаж за 2000 год заимствован из ежегодных отчетов этой фирмы", не содержит достаточно информации, поскольку, например, показатель объема продаж за 2000 год может быть пересмотрен в 2001 году и год или два спустя появиться в годовом отчете уже в виде другого числа.

Ниже приведено несколько примеров наиболее характерных типов ссылок.

1. Если вы *цитируете текст из какой-то другой книги*, ваша ссылка должна включать фамилию автора (авторов), год публикации, название книги, место издания, название издательства, а также номер страницы, содержащей цитируемый материал. Вот пример абзаца с цитатой и подстрочного примечания к этой цитате.

   Чтобы во всех необходимых случаях воздать должное авторам цитируемого материала, а также избежать обвинений в плагиате, в отчете следует указывать соответствующие ссылки. Так, Сигел утверждает:

   "Если вы используете в своем отчете какие-то тексты, данные или идеи из сторонних источников, необходимо сделать ссылку на эти источники. Ссылка указывает на тип материала, заимствованного вами из стороннего источника информации, и позволяет читателю самостоятельно обратиться – при необходимости – к этому источнику".[1]

   [1]Это цитата из книги Siegel A. F., Practical Business Statistics, 4th ed. (New York: Irwin/McGraw-Hill, 2000), p. 549.

2. Если вы *заимствуете из какой-то другой книги идею*, которую объясняете собственными словами (а не в виде дословной цитаты), можно поступать следующим образом.

   Как указывают Bovée и Arens, комбинированные тарифы для газетной рекламы зачастую применяются в случае, когда одно и то же рекламное объявление помещается в двух "родственных" изданиях одной и той же газеты.[1] Это могут быть утренний и вечерний выпуски одной и той же газеты или две разные, но "родственные" газеты.

   [1]Этот материал освещается в книге Bovée C. L. and Arens W. F. Contemporary Advertising, 2nd ed. (Burr Ridge, Ill.: Richard D. Irwin, 1986), p. 413.

3. Если вы *заимствуете данные из какой-то журнальной статьи*, ваша ссылка должна включать название статьи, фамилию автора (если она указана), название журнала, дату и номер страницы. Если эти данные взяты из "третьего" источника, следует указать и его. Например.

   Несмотря на то что эффективность компаний, входящих в список Business Week 50, в среднем снизилась (падение составило 9,5% за период с 20 марта по 18 сентября 1998 г.), некоторые из этих компаний продемонстрировали впечатляющие результаты, например Dell Computer (рост – 85%), Cisco Systems (рост – 44%), EMC (рост – 41%) и Gap (рост – 33%).[1] Разумеется, нет никаких гарантий, что эти замечательные показатели сохранятся и в дальнейшем.

   [1]Данные заимствованы из статьи "The Corporation, Performance: Hey, Things Were Tough All Over", Business Week, 1998, October 19, p. 66. Их источником является Standard & Poor's Compustat, подразделение McGraw-Hill Companies.

4. Если ваши материалы *заимствованы из электронных сетей* (таких как Internet), ваши ссылки должны включать фамилию автора (авторов), название, дату размещения материала в сети (и обновления) — если эти сведения указаны в соответствующем документе, дату вашего обращения к используемому материалу и электронный адрес в виде унифицированного указателя ресурсов (uniform resource locator — URL). Следует предоставить всю необходимую информацию, чтобы в случае изменения или прекращения действия указанного URL-адреса ваши будущие читатели смогли выполнить поиск в сети с целью попытаться найти требуемый материал. Ниже приведен пример абзаца и подстрочного примечания, в котором дата указывается лишь один раз, поскольку обращение к соответствующему материалу выполнялось в день его размещения в сети.

   Хорошие новости для американских компаний, занимающихся розничной торговлей. Эти новости касаются нового торгового сезона, последовавшего за рождественскими праздниками (декабрь 1998 г.): объем продаж в январе оказался достаточно высоким. В частности, рост объемов продажи в одном из магазинов сети Lehman Brothers в январе 1999 г. составил 8,7%, а в сопоставимом с ним по размерам магазине сети Wal-Mart объемы продажи выросли на 10,3%.[1]

   [1]Из G. Crawford, "U.S. Retailers Post Strong January Sales", Yahoo! News, `http://dailynews.yahoo.com/headlines/bs/story.html?s=v/nm/19990204/bs/retailing_2.html`, 1999, February 4.

5. Если речь идет о материале, который вы *почерпнули из интервью, письма или телефонного звонка*, значит, вы должны сделать ссылку на *личный контакт*. При этом следует указать фамилию и должность соответствующего лица, а также место и дату контакта с ним. Например.

> Одно из наиболее важных замечаний, касающихся фьючерсных рынков, заключается в том, что "они дают возможность инвесторам уклоняться от риска с относительно низкими затратами".[1] Это позволяет вам изменять характеристики "риск-прибыль" портфеля в соответствии со своими личными предпочтениями.
>
> [1]Профессор Avraham Kamara, Department of Finance, University of Washington, личный контакт, March 30, 1989.

Если вы хотите получить более подробную информацию о ссылках, обратитесь к справочнику *The Chicago Manual of Style* — превосходному источнику подобных сведений.[3]

Раздел приложений

Приложения содержат весь вспомогательный материал, который, на ваш взгляд, является достаточно важным, чтобы присутствовать в отчете, и, тем не менее, не столь уж значительным (возможно, из-за ограниченного объема отчета), чтобы включать его в основной материал. Приложения могут содержать распечатки исходных данных (если их можно уместить на нескольких страницах) с указанием их источника. Можно также при необходимости включить некоторые подробности проекта вашего исследования, ряд дополнительных графических материалов и таблиц, а также дополнительные технические пояснения утверждений, встречающихся в основном тексте отчета. Для удобства читателей материал можно организовать в виде нескольких подразделов, например *приложение А*, *приложение Б* и т.д.

Использование приложений — еще один способ "облегчить жизнь" читателю. Чрезмерные технические подробности не будут раздражать не слишком подготовленного читателя, а специалист всегда сможет обратиться к интересующим его деталям. Например.

> Поскольку влияние пола работников не было статистически значимым, анализ множественной регрессии был повторно выполнен после исключения переменной "Пол". Результаты не претерпели существенных изменений (подробнее об этом — в приложении С).

13.2. Рекомендации и советы

В этом разделе приведены некоторые рекомендации и советы, которые помогут вам сэкономить время и сделать отчет более эффективным.

Помните о своей аудитории

Краткость — сестра таланта. Помните, что ваши читатели — тоже достаточно занятые люди, к рабочему времени которых следует относиться с уважением. Вы окажете им неоценимую услугу, включив в свой отчет лишь самые важные материалы, результаты, диаграммы и выводы. Если вы не можете обойтись в отчете без множества "технических" материалов, поместите их в приложения.

Выражайте свои мысли простым и понятным языком. Не скупитесь на вводный и "ориентировочный" материал, чтобы подготовить недостаточно искушенных читателей к восприятию основного текста отчета.

Просмотрите несколько раз черновой набросок своего отчета. Попытайтесь поставить себя на место читателей. Постарайтесь забыть, что вы несколько недель жили и дышали этим проектом, и вообразите, будто имеете лишь поверхностное представление о его тематике. Подумайте, в какой мере выполненное вами исследование и отчет об этом исследовании связаны с повседневной жизнью. Постарайтесь предусмотреть в отчете свидетельства такой связи — это поможет активнее включить вашего читателя в суть исследуемой проблемы.

О чем писать в первую очередь, во вторую, в последнюю?

Последовательность изложения материала играет далеко не последнюю роль. Вы не сможете написать статью, пока не получите требуемых результатов. Многие исследователи пишут введение и реферат *в последнюю очередь*, поскольку лишь в самом конце им становится понятно, в какой материал они должны "вводить" своих читателей и что им нужно резюмировать.

Прежде всего, нужно выполнить анализ. Исследуйте данные, просмотрите свои графические материалы, вычислите статистические оценки и выполните проверки гипотез. Возможно, придется даже выполнить множественную регрессию с другими X-переменными. Таким образом, вы получите на своем компьютере папку файлов с таким множеством материалов, которое вряд ли сможете сразу же использовать в полном объеме. Тем не менее все эти материалы следует сохранить — возможно, кое-что из этой информации вам понадобится в дальнейшем.

Далее, выберите из файла с результатами анализа наиболее важные из них. Теперь, когда вам уже известно, как обстоят дела, вы готовы к тому, чтобы составить эскиз (краткий набросок) раздела "Анализ и методы". Возможно, на данном этапе вы будете также в состоянии сделать какие-то выводы.

Все, что вам остается, — это написать на основе предварительных кратких набросков соответствующий подробный текст, решить, что следует поместить в приложения, подготовить ссылки, а также написать введение и реферат. Выполнив этот "эскизный проект", перечитайте его несколько раз (помня об аудитории, на которую он рассчитан) и внесите окончательную правку. Если можно, попросите прочитать его кого-то из своих знакомых. Распечатайте его. Ваш отчет готов!

Другие источники

Есть немало источников информации, касающейся написания различных документов, в которых освещаются вопросы стиля и языка изложения материала. Укажем лишь некоторые из этих источников.

1. Если вы не уверены в правильности написания или применения того или иного слова, пользуйтесь *словарями*. Для проверки правописания (орфографии) очень удобны компьютерные текстовые редакторы.
2. Для поиска синонимов (слов одинакового смыслового значения) пользуйтесь *тезаурусами*. Они помогут вам подобрать самое подходящее слово и

[3] См. "Documentation 1: Notes and Bibliographies" и "Documentation 2: Author-Date Citations and Reference Lists". *The Chicago Manual of Style*, 14th ed. (Chicago: University of Chicago Press, 1993), Chapters 15, 16.

избежать чрезмерного повторения одних и тех же слов. Многие компьютерные текстовые редакторы содержат встроенные тезаурусы.

3. Существует немало книг, предназначенных для менеджеров, которым приходится время от времени составлять отчеты, включающие те или иные технические материалы. В этих книгах можно найти много полезных рекомендаций. Приведем лишь несколько из них: *Lesikar* and *Pettit, Report Writing for Business, Tichy* and *Fourdrinier, Effective Writing for Engineers, Managers, Scientists* и *Weisman, Basic Technical Reporting.*[4]
4. Более подробно о правилах написания текстов можно прочитать в *The Chicago Manual of Style.*

13.3. Пример: формула оперативного ценообразования для ответа на запросы потребителей

Ниже приведен пример отчета, основанного на анализе множественной регрессии. Этот отчет соответствует организационному плану отчета, представленному ранее в этой главе. Обратите внимание: основной акцент в отчете сделан на практическом применении полученных результатов, а не на технических подробностях исследования.

Формула оперативного ценообразования для ответа на запросы потребителей

Составлен для
Бонни С. Веннерстрома, вице-президента по сбыту

Составлен
Кларой Х. Сигел, директором по исследованиям
Mount Olympus Design and Development Corporation
10 апреля 1999 г.

Реферат

Мы теряем потенциальных клиентов, поскольку не в состоянии оперативно устанавливать цены на нашу продукцию. Наши торговые представители жалуются, что к тому моменту, когда мы на следующий день сообщаем им цену, многие из этих клиентов уже успевают заключить соглашения с каким-либо из наших конкурентов. Наше предложение заключается в создании некоей "формулы оперативного ценообразования". В этом случае потенциальные покупатели стандартной продукции смогут узнавать ее приблизительную цену по телефону. Это поможет поддерживать их заинтересованность в нашей фирме, пока они будут ожидать точную цену, которая будет сообщена на следующий день. Тем самым мы будем не только напоминать клиентам о своем существовании, но и более оперативно реагировать на их потребности.

[4] Lesikar R. V. and Pettit J. D. *Report Writing for Business*, 10th ed. (New York: Irwin/McGraw-Hill, 1998); Tichy H. J. and Fourdrinier S. *Effective Writing for Engineers, Managers, Scientists*, 2nd ed. (New York: Wiley, 1988); Weisman H. M. *Basic Technical Reporting* (Englewood Cliffs, NJ: Prentice-Hall).

Введение

Для определения точной цены в настоящее время требуется от трех до шести часов инженерных работ. Когда наши клиенты хотят сделать заказ на новое изделие, они хотят знать диапазон цен, что даст им возможность сравнивать продукцию конкурирующих фирм. В прошлом, когда число наших конкурентов было невелико, это не представляло особой проблемы. Но в последнее время, несмотря на превосходное качество нашей продукции, все большее число заказов перехватывают наши конкуренты, которые предоставляют потенциальным клиентам более оперативную информацию.

После консультаций с технологическим отделом мы пришли к выводу о том, что временная задержка неизбежна, если речь идет об определении точной цены. Чтобы определить точные размеры конструкции и требования к электропитанию (что, в свою очередь, определяет величину наших затрат), требуется определенный объем предварительной, прикидочной работы.

Кроме того, мы проконсультировались с несколькими ведущими клиентами. Несмотря на то что они не настаивают на немедленном предоставлении точной цены, было бы неплохо, если бы мы могли в ходе первоначального контакта с нашим представителем давать им хотя бы приблизительное представление о цене. Тем самым мы удовлетворили бы по крайней мере два из их требований: (1) они получили бы определенное свидетельство нашей конкурентоспособности и (2) эта ценовая информация помогла бы им в выработке собственных конструкторско-технологических решений, поскольку они смогли бы быстро оценить несколько различных походов.

Основываясь на нашем собственном опыте, мы разработали формулу, которая позволяет чрезвычайно быстро получить приблизительную величину наших затрат.

Оперативная оценка затрат = $1 356 + $35,58(компоненты) + $5,68(размеры).

Результирующая "оперативная оценка затрат", скорее всего, будет отличаться от подробной калькуляции расходов не более чем на $200. Оперативную информацию о цене можно сообщить по телефону, предварительно добавив соответствующую (конфиденциальную) наценку, зависящую от дисконтного класса данного клиента.

Мы собрали данные о точных расчетах цен за последнее время. Эти данные взяты из наших внутренних компьютерных записей, соответствующих детальным расценкам, которых мы обычно придерживаемся в течение семидневного периода. Мы проанализировали следующие важнейшие переменные.

1. **Затраты**, подсчитанные технологическим отделом. Это внутрифирменный конфиденциальный материал. Данная переменная подлежит прогнозированию. Это единственная переменная, не доступная на стадии начальных телефонных переговоров.
2. **Количество компонентов**, используемых в конструкции. Это примерный показатель сложности конструкции; он почти всегда указывается клиентом на стадии начальных переговоров.
3. **Размеры конструкции**. Это весьма приблизительный показатель фактического размера конечного изделия. Он указывается клиентом в качестве "отправной точки".

Из 72 расценок, выданных за этот квартал, мы отобрали 56 как репрезентативные с точки зрения стандартных работ, с которыми нам приходится встречаться чаще всего. Случаи, которые мы отбросили, либо требовали применения специального химического процесса (или покрытия), либо использовали необычные компоненты, постоянных поставщиков которых у нас нет. Полученная совокупность данных приведена в приложении.

Анализ и методы

Этот раздел начинается с описания наших типичных расценок и завершается представлением формулы прогнозирования затрат (использована методология множественной регрессии) и ее интерпретацией.

Вот профиль наших самых типичных работ. Из гистограмм, показанных на рис. 1, нетрудно заметить, что наша типичная цена включает затраты от $3 000 до $5 000, редко больше или меньше. Стандартное отклонение затрат равняется $707, указывая примерную величину ошибки, которую мы совершили бы, если бы (по глупости!) предлагали оперативные расценки, основываясь исключительно на средней величине затрат — $3 987. Количество компонентов, как правило, составляет от 10 до 50 (более сложные конструкции встре-

чаются редко). Размеры конструкции обычно находятся в диапазоне от 200 до 300. Проблем с резко отклоняющимися значениями нет, поскольку все крупные (или нетипичные) конструкции рассматриваются отдельно как особые случаи, к которым наша формула оперативной оценки затрат неприменима.

Далее, мы рассмотрели взаимосвязь между затратами и каждой из остальных переменных. Как видно из двух диаграмм рассеяния, показанных на рис. 2, наблюдается очень сильная взаимосвязь между количеством компонентов и нашими затратами (соответствующий коэффициент корреляции равен 0,949) и сильная взаимосвязь между размерами конструкции и нашими затратами (соответствующий коэффициент корреляции составляет 0,857). Эти сильные взаимосвязи указывают на то, что мы действительно сможем получать достоверные прогнозы затрат, основываясь на этих переменных. Взаимосвязь между количеством компонентов и размерами конструкции — умеренная (коэффициент корреляции равен 0,760; соответствующая диаграмма рассеяния приведена в приложении). Поскольку эта взаимосвязь не является идеальной, размеры конструкции могут привносить в общую картину полезную дополнительную информацию. Кроме того, указанные взаимосвязи являются линейными, что свидетельствует о возможности выполнения регрессионного анализа.

Анализ множественной регрессии для прогнозирования затрат на основании других переменных (количества компонентов и размеров конструкции) позволил получить следующее уравнение прогнозирования:

прогнозируемые затраты = $1 356 + $35,58(компоненты) + $5,68(размеры).

Эти прогнозируемые затраты можно легко вычислить, исходя из информации, сообщаемой клиентом по телефону. Они представляют наш наиболее достоверный прогноз величины детальных затрат, полученный методом наименьших квадратов на основании линейной модели этого типа. Это и есть "оперативная оценка затрат", которую мы предлагаем.

Это уравнение прогнозирования оказывается вполне разумным. Оценочная величина фиксированных затрат, составляющая $1 356, с лихвой перекрывает наши обычные накладные расходы. Оценочная величина затрат на один компонент, составляющая $35,58, несколько больше того, на что можно было бы рассчитывать, поскольку оптимальный прогноз включает другие факторы (такие как заработная плата работников), которые также возрастают с увеличением сложности конструкции. Сумма $5,68 на едини-

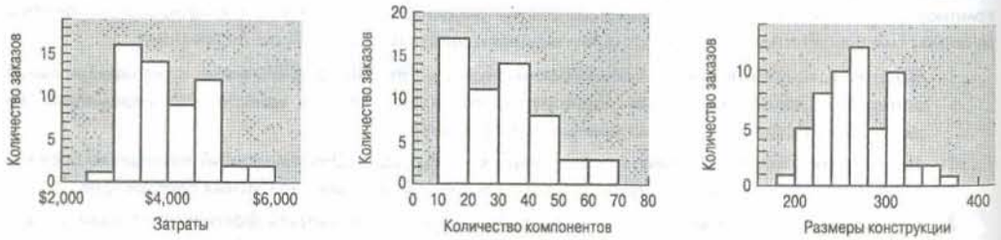


Рис. 1. Гистограммы переменных

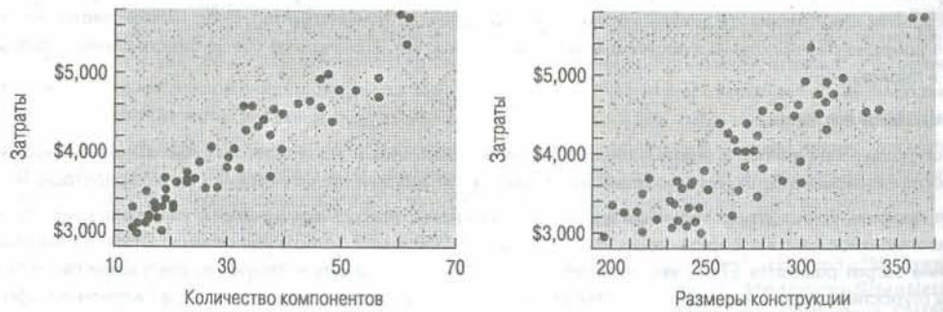


Рис. 2. Диаграммы рассеяния для переменных

цу размеров конструкции опять-таки выше, чем наши фактические затраты, поскольку величина размеров конструкции также представляет информацию о других дорогостоящих аспектах конструкции.

Ниже приведен пример использования предлагаемого уравнения прогнозирования. Рассмотрим клиента, который намерен разместить заказ на изготовление конструкции, включающей 42 компонента; размер этой конструкции равняется примерно 315. Наши затраты можно оценить следующим образом:

прогнозируемые затраты = $1 356 + $35,58 × 42 + $5,68 × 315 = $4 640.

Если к этому клиенту обычно применяется 20% наценка (в соответствии с конфиденциальными записями, которые легко можно получить на компьютере), тогда расценку можно определить следующим образом:

оперативная расценка = $4 640 × 1,2 = $5 568.

Насколько точны эти оперативные расценки? Обычно они не должны отклоняться более чем на $200 от детальной калькуляции затрат (которая на начальном этапе отсутствует), если исходить из стандартной ошибки оценки, равной $169.[5]

Существуют три подхода, которыми можно было бы воспользоваться для учета этой остающейся ошибки. Во-первых, мы могли бы сказать клиенту, что это число является довольно приблизительным и что окончательная сумма будет определяться в результате обычной детальной калькуляции затрат. Во-вторых, — рассмотрим другую крайность, — мы могли бы сразу же сообщить клиенту окончательную сумму; для этого нам, возможно, пришлось бы добавить несколько сотен долларов в качестве "запаса прочности". Наконец, мы могли бы оставить за собой право пересмотреть указанную сумму, гарантируя при этом клиенту, что окончательная цена не поднимется выше некоторой границы (например, $100).

Насколько точно позволяют прогнозировать затраты такие показатели, как количество компонентов и размеры конструкции? Рассмотрим всю вариацию затрат для различных заказов: очень большая часть этой вариации объясняется количеством компонентов и размерами конструкции ("R-квадрат" равен 94,5%). Это вряд ли можно объяснить случайностью, поскольку уравнение регрессии представляется очень высоко статистически значимым.[6]

Средняя величина затрат составляет $3 987. Используя вместо этой средней величины затрат предлагаемое уравнение прогнозирования, мы снижаем размер нашей ошибки с $707 (обычное стандартное отклонение затрат) до $169 (стандартная ошибка оценки из регрессионного анализа).

Действительно ли для прогнозирования затрат нам требуется как количество компонентов, так и размеры конструкции? Да, поскольку дополнительный вклад каждой из этих переменных (помимо информации, обеспечиваемой другой переменной) является очень высоко статистически значимым в соответствии с *t*-тестом каждого из коэффициентов регрессии.

Мы также попытались спрогнозировать наличие возможных технических проблем, но не выявили их. Например, диагностическая диаграмма в приложении демонстрирует отсутствие дополнительных структур, которые можно было бы использовать для дальнейшего улучшения результатов.

Выводы и резюме

Мы в состоянии обеспечить нашим клиентам более качественное обслуживание, предоставляя им немедленно расценки для наиболее типичных заказов на основании следующего уравнения прогнозирования затрат:

прогнозируемые затраты = $1 356 + $35,58(компоненты) + $5,68(размеры).

---

[5] Для обычной линейной модели можно ожидать, что примерно 2/3 оперативных показателей затрат для этой совокупности данных будут отклоняться не более чем на $169 от соответствующих им подробных точных калькуляций затрат. При выполнении аналогичных заказов в будущем эта ошибка может быть несколько выше вследствие неопределенности коэффициентов регрессии, входящих в уравнение прогнозирования. О величине ошибки для *других* заказов в будущем трудно сказать что-либо определенное.

[6] Соответствующее p-значение меньше 0,001 и указывает вероятность найти столь сильную прогнозирующую взаимосвязь в том случае, если бы на самом деле взаимосвязь отсутствовала и имела место лишь чистая случайность.

Добавляя наценку и, возможно, несколько сотен долларов для компенсации ошибки прогнозирования, мы могли бы немедленно предоставлять своим клиентам расценки в нескольких различных формах.

Сообщать расценку лишь в виде *приблизительной* цены. Сказать клиенту, что фактическая цена будет определяться в результате обычной детальной калькуляции затрат и будет известна на следующий день. В сущности, весь риск неопределенности цены перекладывается на клиента.

Сообщать твердую расценку. Для этого придется добавить небольшую дополнительную сумму и перенести таким образом риск неопределенности цены с клиента на нашу фирму.

Компромиссный вариант. Сообщать *примерную* котировку, но ограничить риск клиента, зафиксировав возможное изменение цены. Например, мы могли бы оставить за собой право пересмотреть указанную сумму, гарантируя при этом клиенту, что окончательная цена не поднимется выше некоторой границы (скажем, $100). Следует также рассмотреть возможность *снижения* цены, если предварительный ее прогноз оказался завышенным.

Это уравнение прогнозирования затрат основывается на нашем фактическом производственном опыте и на хорошо известных статистических методах. Анализ множественной регрессии подходит для большинства типичных заказов, с которыми приходится иметь дело нашей фирме; полученные результаты являются очень высоко статистически значимыми.

Если мы внедрим у себя в фирме метод "оперативных расценок", то должны помнить о следующей "проблеме выбора": когда клиенты начнут догадываться о том, как мы устанавливаем свои расценки, они могут попытаться искусственно раздувать сложность своих заказов. Поскольку наше уравнение прогнозирования основывается на самых *типичных* заказах, серьезное изменение в уровне сложности заказа может привести к неправильному ценообразованию. В ответ мы могли бы выявить источник этой дополнительной сложности и соответственно время от времени модифицировать модель оперативного расчета цен.

Если результаты применения предлагаемого метода окажутся удачными, можно было бы подумать о расширении этой программы с целью предоставления оперативных расценок по дополнительным категориям проектно-конструкторских работ.

Источники информации

Используемые в настоящем отчете данные были получены из конфиденциальных корпоративных записей в системе по состоянию на 4/6/99.

В качестве статистического программного обеспечения использовался пакет StatPad, торговая марка Skyline Technologies, Inc.

Пояснение общих статистических принципов применительно к сфере бизнеса содержится в книге Siegel A. F. Practical Business Statistics, 4th ed. (New York: Irwin/McGraw-Hill, 2000), перевод которой вы держите в руках.

Приложение

Ниже приведены данные, которые подвергались анализу. В эту совокупность данных включены только стандартные проекты. Шестнадцать проектов, в которых либо требовалось применение специального химического процесса (или покрытия), либо использовались необычные компоненты, для снабжения которыми у нас нет постоянных поставщиков, были исключены из рассмотрения.

Количество компонентов	Размеры конструкции	Затраты, дол.	Количество компонентов	Размеры конструкции	Затраты, дол.
27	268	4 064	42	288	4 630
23	243	3 638	24	244	3 659
30	301	3 933	37	220	3 712
16	245	3 168	23	235	3 677
37	265	4 227	32	250	3 826

Количество компонентов	Размеры конструкции	Затраты, дол.	Количество компонентов	Размеры конструкции	Затраты, дол.
14	233	3 105	28	267	3 576
20	247	3 352	38	309	4 547
33	334	4 581	49	317	4 806
23	290	3 708	56	313	4 717
35	314	4 360	14	196	2 981
31	270	4 058	46	314	4 949
17	236	3 162	13	248	3 032
47	322	5 008	39	274	4 051
52	309	4 790	33	262	4 283
34	341	4 593	17	264	3 250
25	271	3 869	44	277	4 280
26	252	3 572	44	299	4 665
19	300	3 677	17	233	3 389
16	224	3 211	17	206	3 296
30	280	3 840	61	358	5 732
36	257	4 428	56	302	4 963
61	306	5 382	15	241	3 109
19	216	3 518	48	271	4 419
16	277	3 496	39	297	4 524
46	280	4 588	19	232	3 425
60	366	5 752	20	201	3 368
18	217	3 042	13	213	3 295
18	242	3 358	21	237	3 605

Ниже приведена компьютерная распечатка анализа множественной регрессии.

Коэффициенты корреляции:

	Затраты	Компоненты	Размеры
Затраты	1	0,949313	0,856855
Компоненты	0,949313	1	0,759615
Размеры	0,856855	0,759615	1

Уравнение регрессии:

затраты = 1356,148 + 35,57949 * (компоненты) + 5,678224 * (размеры)

S = 169,1943

R2 = 0,944757

Статистический вывод для затрат на уровне 5%:

Уравнение регрессии действительно объясняет значимую долю вариации затрат.

F = 453,2050 с 2 и 53 степенями свободы

Переменная	Влияние на затраты	95% доверительный интервал		Проверка гипотез	Стандартная ошибка коэффициента	t-статистика
	Коэффициент	от	до	Значимый?	Стандартная ошибка	t
Константа	1356,148	983,1083	1729,189	Да	185,9845	7,291728
Компоненты	35,57949	30,55847	40,60051	Да	2,503300	14,21303
Размеры	5,678224	3,916505	7,439943	Да	0,878329	6,464801

На рис. 3 показана диаграмма рассеяния для двух независимых (объясняющих) переменных: количества компонентов и размеров конструкции. Коэффициент корреляции равен 0,760.

Диагностическая диаграмма ошибок прогнозирования затрат в зависимости от прогнозируемых затрат (рис. 4) демонстрирует отсутствие какой-либо структуры — лишь случайный разброс точек данных. Это свидетельствует об отсутствии простых способов улучшения качества прогнозирования затрат на основании количества компонентов и размеров конструкции.

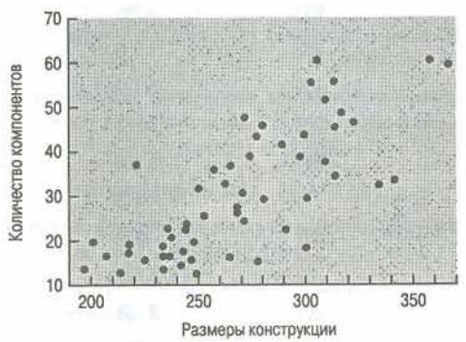


Рис. 3. Диаграмма рассеяния количества компонентов в зависимости от размеров конструкции

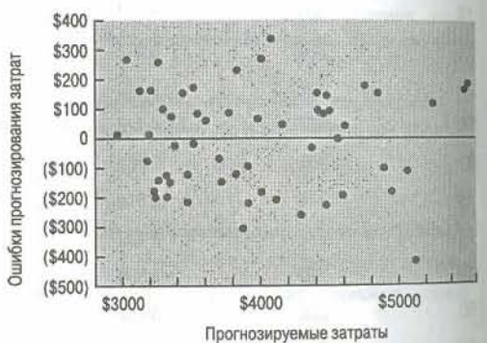


Рис. 4. Эта диагностическая диаграмма свидетельствует об отсутствии каких-либо очевидных проблем

13.4. Дополнительный материал

Резюме

Изложение того или иного материала, который требуется довести до сведения определенной аудитории, является важной составляющей профессии менеджера. Статистические отчеты помогут вам в понятной и объективной форме изложить базовые сведения, касающиеся исследуемой ситуации. Будьте снисходительны к своим читателям и постарайтесь объяснить все полученные вами результаты на понятном им языке. Четко сформулированная *цель* поможет вам ограничить тематику отчета и сосредоточить внимание лишь на важнейших и непосредственно относящихся к делу вопросах. Четкое определение своей *аудитории* поможет вам выбрать подходящий стиль изложения материала и уровень его детализации.

Оптимально организованный отчет характеризуется четкостью и простотой. Безусловно, автор отчета хотел бы, чтобы при ознакомлении с материалом у его читателей возникало как можно меньше проблем. Прежде всего необходимо составить план, в котором, возможно, каждый абзац будущего отчета представлен одной строкой. Ниже приведен примерный план организации отчета.

1. **Реферат** представляет собой абзац в самом начале отчета, где описываются наиболее важные факты и выводы из проделанной вами работы.
2. **Введение** состоит из нескольких абзацев, в которых вы описываете "предысторию вопроса", цель исследования и данные, с которыми вы работали.
3. Раздел **Анализ и методы** позволяет вам интерпретировать исходные данные, включив в отчет их графическое представление, итоговые статистические показатели и результаты, которые вы объясняете по ходу дела.
4. Цель раздела **Выводы и резюме** заключается в том, чтобы представить общую картину ситуации, собрав воедино все важнейшие мысли, которые вы хотели бы довести до сведения своей аудитории.
5. Раздел **Ссылки** содержит указания на сведения, полученные вами из внешних источников. Назначение этого раздела — предоставить возможность своим читателям в случае необходимости самостоятельно обратиться к этим источникам. Эти ссылки можно указывать по ходу текста (на соответствующих страницах) или собрать их в самостоятельный раздел.
6. **Приложения** должны содержать весь вспомогательный материал, который, на ваш взгляд, является достаточно важным, чтобы присутствовать в отчете и, тем не менее, не столь уж значительным, чтобы включать его в основной текст.

Кроме того, ваш отчет может содержать *титульную страницу* и *оглавление*. **Титульная страница** идет первой и включает название отчета, фамилию и должность лица, для которого вы подготовили этот отчет, вашу фамилию и должность (как основного исполнителя) и дату. **Оглавление** следует после реферата и отражает структуру отчета (с указанием номеров страниц).

Прежде всего следует выполнить анализ, затем отобрать из своего файла анализа наиболее важные результаты. После этого необходимо составить план раздела "Анализ и методы" и подготовить выводы. Текст абзацев этого раздела создается на основе пунктов предварительно составленного вами плана. Затем нужно решить, что включать в приложение, подготовить ссылки, написать введение и реферат.

Подготовленный вами материал должен быть четким и лаконичным. Перечитайте свой отчет еще раз, представляя себе аудиторию, для которой он предназначен.

Основные термины

- Титульная страница (title page), 723
- Оглавление (table of contents), 723
- Реферат (executive summary), 723
- Анализ и методы (analysis and methods), 724
- Выводы и резюме (conclusion and summary), 726
- Ссылки (reference), 726
- Приложение (appendix), 728

Контрольные вопросы

1. Какова главная цель написания отчета?
2. Почему необходимо определить цель отчета и аудиторию, для которой он предназначен?
3. Как может способствовать написанию отчета предварительное составление его плана?
4. Приведите доводы в пользу включения в свой отчет статистических результатов.
5. Следует ли изъять из реферата основные результаты, завершив его примерно такой фразой: "Мы исследовали эти вопросы и можем предложить вашему вниманию некоторые рекомендации"? Поясните свой ответ.
6. Как можно использовать реферат и введение для того, чтобы с отчетом могли познакомиться люди, очень ограниченные во времени?
7. Можно ли повторять во введении тот материал, который уже использовался в реферате?
8. а) Какого рода материал появляется в разделе "Анализ и методы"?

   б) Следует ли описывать все, что вы исследовали, в разделе "Анализ и методы"? Поясните свой ответ.
9. Следует ли полагать, что каждому, кто знакомится с вашими выводами, уже известны все подробности раздела "Анализ и методы"?
10. а) Укажите две причины, по которым в отчет следует включить ссылки, если вы используете в нем материалы из Internet, из какой-либо книги, журнала или иного источника.

    б) Что позволяет вам с уверенностью говорить о том, что ваша ссылка содержит всю необходимую информацию?

    в) Как вы организуете ссылку на материал, позаимствованный вами из телефонного разговора или интервью со специалистом?
11. а) Какой материал следует помещать в приложение?

    б) Как приложение помогает удовлетворить интересы не только "случайного", но и "целенаправленного" читателя?

12. Как вы можете помочь тем читателям вашего отчета, которые испытывают нехватку времени?
13. Когда лучше всего приступать к написанию введения и реферата — в самом начале или в последнюю очередь?
14. Какова взаимосвязь между планом отчета и его окончательным вариантом?
15. Как можно убедиться в правильности употребления какого-то слова?
16. Как отыскать синонимы для данного слова? Зачем это может понадобиться?

Задачи

1. Ваш начальник попросил вас составить отчет. Определите цель и аудиторию отчета в каждом из перечисленных ниже случаев.

   а) Фирма пытается расширить сферу своих поставок. Необходим базовый материал, касающийся производственных мощностей других фирм.

   б) Новая стереосистема практически готова к поставке потребителям. Она, несомненно, превосходит по своим характеристикам все существующие системы на рынке. Ваш начальник обратился в редакцию журнала, посвященного высококачественной электронике, с просьбой включить в рубрику "Новая продукция" материал о вашей стереосистеме.

   в) В последнее время довольно часто происходят поломки производственного оборудования, точную причину которых никому не удается установить. К счастью, вы располагаете информацией, касающейся подробностей каждой такой поломки.

   г) Банк вашей фирмы отказывается от дальнейшего предоставления кредитов, утверждая, что объемы продаж в вашей промышленной группе слишком тесно привязаны к состоянию экономики и, следовательно, слишком подвержены влиянию финансовых проблем, проявляющихся во время экономического спада. Ваш начальник полагает, что реальные данные, характеризующие объем продаж вашей фирмы и валовой национальный продукт США, могут свидетельствовать об обратном.
2. Заполните пробелы, вставив слова *влиять* или *определять*.

   а) Объемы сверхурочных работ и время суток значительно _____ на уровень травматизма на производстве.

   б) Объемы сверхурочных работ и время суток в значительной мере _____ на уровень травматизма на производстве.

   в) Интересно отметить, что величина стажа работы, по-видимому, не _____ на производительность этих работников.

   г) Интересно отметить, что величина рабочего стажа, по-видимому, не _____ на производительность этих работников.
3. Напрягите свое воображение и подготовьте текст реферата некоторого гипотетического отчета, взяв за основу каждую из ситуаций, описанных в задаче 1. (*Примечание.* Для каждой из этих ситуаций будет достаточно одного абзаца.)

4. Для каждого из следующих предложений скажите, в какой раздел (или разделы) отчета его можно включить.

   а) Диаграмма рассеяния дефектов в зависимости от объема продукции демонстрирует умеренно сильную взаимосвязь (коэффициент корреляции равен 0,782), указывая на то, что наши самые сложные проблемы возникают в периоды наивысшего спроса — именно тогда, когда их появление совершенно недопустимо.

   б) Третий вариант — продажу нашего подразделения независимой сторонней компании — следует рассматривать как крайнюю меру, к которой можно прибегнуть лишь в том случае, если две другие возможности не приведут к желаемому результату.

   в) Эти проблемы начали проявляться еще пять лет тому назад, когда бы введен в строй новый завод, и обошлись нашей фирме за последние два года по меньшей мере в $2 000 000.

   г) Ниже приведена совокупность данных, отражающая цены каждого изделия на каждом из исследованных нами рынков.

5. Переупорядочьте информацию так, чтобы в каждом из перечисленных ниже случаев получилась правильная ссылка.

   а) Название статьи из *The Wall Street Journal*: "Может ли руководитель финансового отдела вытащить кролика из шапки?" Она опубликована 16 марта 1989 г., т.е. в четверг. Статья помещена на странице C1. Она написана Джеймсом А. Уайтом, штатным корреспондентом. В статье говорится о том, как правильный выбор метода оценки позволяет руководителю финансового отдела представить свою деятельность в наиболее выгодном для себя свете.

   б) Чтобы удостовериться в правильности некоторых технических подробностей отчета об аренде и налогах, вы решили позвонить эксперту в Вашингтонском университете. Профессор Лоуренс Д. Шолль подтверждает вашу догадку о том, что законодательство в этой сфере является чересчур сложным, и высказывает предположение, что данная проблема вообще не имеет простых решений.

   в) Вы воспользовались книгой *Управление качеством: инструменты и методы усовершенствования* в качестве базового материала для отчета, посвященного проблемам совершенствования производства (авторы книги Говард Гитлоу, Алан и Роза Оппенхайм). Наибольшее внимание вы уделили главе 8. Книга была опубликована в 1995 г. издательством Richard D. Irwin, Inc., Бериджж, шт. Иллинойс. Книга защищена авторским правом; ее ISBN — 0-256-10665-7. Она посвящена "бесконечному совершенствованию людей".

   г) Получена информация о том, что доверие потребителей достигло высшей отметки — 127,6. Эта информация появилась 26 января 1999 г. Вы обнаружили ее в Internet 4 февраля 1999 г. на странице, принадлежащей The Conference Board и озаглавленной "Показатели делового цикла, последние выпуски".
   URL-адрес: `http://www.tcb-indicators.org/`.

6. Какая важная информация упущена в каждой из приведенных ниже ссылок?

   а) Личные контакты, 1998.

   б) *Business Week*, p. 80.

   в) *Basic Business Communication* (Burr Ridge, Ill.: Richard D. Irwin).

   г) White James A. "Will the Real S&P 500 Please Stand Up? Investment Firms Disagree on Index," *The Wall Street Journal*.

   д) Данные были получены из White House Economic Statistics Briefing Room в Internet.

Упражнение с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении А.

Напишите трех- или пятистраничный отчет о взаимосвязи между полом и уровнем заработной платы этих служащих. Обязательно обсудите результаты следующих статистических анализов: (a) *t*-теста (для двух выборок) сравнения заработной платы мужчин и женщин и (b) множественной регрессии, объясняющей заработную плату исходя из возраста, стажа работы и индикаторной переменной, обозначающей пол служащего.

Проект

Выполните анализ множественной регрессии на основе экономических данных по собственному выбору (их источником может быть Internet, библиотека или ваша компания). Подытожьте результаты этого анализа в виде отчета для высшего руководства компании (это может быть обычный отчет, предназначенный "для принятия к сведению", или предложение конкретных действий). Ваш отчет должен состоять из 5–7 страниц с приложением и иметь следующую структуру:

1. *Введение*. Перечислите основные положения, касающиеся рассматриваемой проблемы, исследуемые вами вопросы и рассматриваемую совокупность данных настолько подробно, чтобы образованному человеку, которому неизвестны детали изучаемой ситуации, все стало ясно
2. *Анализ и методы*. Проанализируйте данные, представив соответствующие графические материалы и результаты вычислений. По ходу изложения давайте необходимые пояснения. Возможно, при этом вам придется проделать следующую работу.

   а) *Исследовать* свои данные, построив для каждой переменной гистограмму или блочную диаграмму из прямоугольников, а также диаграмму рассеяния для каждой пары переменных.

   б) Использовать *преобразование* (например, логарифмическое) *лишь в том случае*, если оно действительно помогает проведению анализа (в частности, когда вы сталкиваетесь с какой-либо серьезной проблемой в диагностической диаграмме).

   в) Вычислить *корреляцию* для каждой пары переменных и интерпретировать полученные значения.

г) Представить *множественную линейную регрессию* для прогнозирования одной переменной (выбранной соответствующим образом) на основании других переменных. Прокомментируйте качество этого регрессионного анализа с точки зрения точности прогнозирования (стандартной ошибки оценки) и того, насколько успешно он объясняет соответствующую взаимосвязь (коэффициент детерминации). В частности, постарайтесь указать, насколько логичны и обоснованны полученные вами результаты и являются ли они статистически значимыми.

3. *Выводы и резюме*. О чем свидетельствует проведенный вами анализ с точки зрения исследования интересующей вас ситуации? В какой мере вам удалось ответить на поставленные вопросы?
4. *Приложение*. Перечислите данные, указав их источники. Можно также приложить ксерокопию использованных вами данных. (На количество страниц в отчете это не влияет.)

ГЛАВА **14**

Временные ряды: анализ изменений во времени

Временные ряды отличаются от данных об одном временном срезе в том отношении, что в случае временных рядов сама *последовательность наблюдений несет в себе важную информацию*. В частности, чтобы охарактеризовать какую-либо совокупность данных в целом, вам уже недостаточно знать лишь типичное значение этих данных (например, среднее значение) или даже изменчивость этой совокупности данных (описываемую, например, стандартным отклонением). В этом случае желательно знать, *что, скорее всего, произойдет дальше*. Подобный прогноз должен по возможности точнее экстраполировать ближайшее поведение системы с точки зрения моделей поведения этой системы в прошлом. Ниже приведено несколько примеров ситуаций, описываемых временными рядами.



Первый. Чтобы составить бюджет на следующий квартал, вам требуется достоверная оценка ожидаемого объема продаж. Этот прогноз послужит основой для прогнозирования других показателей бюджета (возможно, с помощью регрессионного анализа). Проанализировав временной ряд фактических квартальных объемов продажи за последние несколько лет, можно выдать прогноз, который будет представлять собой наиболее достоверную оценку, базирующуюся на общих тенденциях продаж (надеемся, положительных), с учетом любых сезонных колебаний спроса. Если, например, всегда наблюдается спад деловой активности по окончании четвертого квартала (который включает рождественские распродажи) и в начале первого квартала, то желательно, чтобы ваш прогноз отражал эту типичную сезонную модель колебаний спроса.

Второй. Чтобы решить, стоит ли строить новый завод, необходимо знать темпы роста вашего целевого рынка. Анализ имеющихся у вас в распоря-

жении данных временного ряда, отражающего объемы продаж и цены в соответствующей отрасли, поможет вам правильно оценить ваши шансы на успех. Впрочем, было бы наивным ожидать точного ответа. Прогнозирование будущего — занятие, связанное с немалым риском и неопределенностью, даже при наличии самых совершенных компьютеров и программного обеспечения. Несмотря на то что анализ временного ряда действительно поможет вам в принятии решений, позволяя сверять их с реально происходящими процессами, немалая доля риска остается и в этом случае.

Третий. Непрерывно отслеживая временные ряды данных о вашей фирме — как внутренних данных (объемы продаж, затраты и т.п.), так и внешних (объемы продаж и импорта в рамках всей отрасли и т.п.), — вы сможете повысить эффективность управления своей фирмой. Экстраполируя на будущее тенденции, выявленные на самых ранних стадиях, вы сможете принимать активное участие в наиболее перспективных и быстроразвивающихся областях и, наоборот, своевременно покидать бесперспективные и приходящие в упадок рынки. Прогнозируя сезонные потребности в наличных деньгах, вы сможете избежать паники ситуаций, связанных с острой нехваткой наличности, а также ситуаций, связанных с ее избытком (что также имеет свои отрицательные стороны). Прогнозируя потребности в товарно-материальных запасах, вы сможете минимизировать потери, связанные с невыполнением заказов (что будет способствовать повышению вашей конкурентоспособности), и затраты (проценты за кредиты и оплата хранения), связанные с поддержанием чрезмерных запасов. Короче говоря, в этих совокупностях данных, относящихся к категории временных рядов, содержится очень много полезной информации.

14.1. Обзор анализа временных рядов

Чтобы методами, описанными в предыдущих главах (например, доверительные интервалы и проверки гипотез), можно было воспользоваться и в случае временных рядов, их необходимо определенным образом трансформировать. Почему? Потому что необходимые условия применения этих методов в данном случае не выполняются. В частности, временной ряд не является случайной выборкой из некоторой генеральной совокупности.[1] Гораздо вероятнее, например, что завтрашняя цена окажется ближе к сегодняшней, чем к прошлогодней цене; последовательные наблюдения *не* являются независимыми друг от друга. Если же, тем не менее, мы вычислим доверительные интервалы и выполним проверки гипотез обычным способом, то возникнет опасность, что наш уровень ошибки окажется намного большим, чем заявленные нами 5%. Анализ временных рядов требует применения особых методов, которые учитывают существование определенной зависимости между наблюдениями. Базовые идеи и концепции статистических выводов остаются теми же, однако методы адаптируются к новой ситуации.

Главная цель анализа временных рядов заключается в прогнозировании будущего. Для описания временных рядов требуется определенная модель. **Модель** (ее еще называют **математической моделью**, или **процессом**) представляет собой систему уравнений, которая позволяет получить некий набор искусственных данных в форме временных рядов. Ниже описана процедура, связанная с прогнозированием.

1. Выберите семейство моделей временных рядов.
2. Оцените конкретную модель (в рамках выбранного вами семейства), которая позволяет получить искусственные данные, отвечающие важнейшим характеристикам (но не каким-то аномалиям и исключениям) анализируемого временного ряда.
3. Ваш **прогноз** будет представлять собой ожидаемое (т.е. среднее) значение будущего поведения модели, для которой сделана оценка. Обратите внимание, что вы можете прогнозировать будущее для той или иной математической модели с помощью компьютера, хотя будущее анализируемого ряда не известно.
4. **Границами прогноза** являются доверительные интервалы для вашего прогноза (если данная модель позволяет определять их); если используемая вами модель корректна, то будущее наблюдение с вероятностью, например, 95% попадет в эти границы. Границы прогноза вычисляются обычным способом на основании стандартной ошибки, которая представляет изменчивость будущего поведения оцениваемой модели.

Следуя этой процедуре, вы не просто делаете прогнозы. Выбрав подходящую модель, которая позволяет получать совокупности данных, "похожие" на ваши фактические ряды, вы глубже уясняете сценарии поведения этих рядов. Этот тип углубленного статистического понимания устройства окружающего мира принесет вам пользу в качестве базовой информации для принятия решений.

Несмотря на то что всем нам требуются достоверные прогнозы будущего, было бы чересчур опрометчивым рассчитывать на стопроцентную точность таких прогнозов. Та точность прогнозов, которую мы хотели бы получить, практически недостижима, поскольку истинно неожиданное событие невозможно предсказать по определению.[2] Однако потребность в прогнозах столь велика, что люди готовы прибегнуть к любым средствам, которые *могут* привести хотя бы к незначительному улучшению. Чтобы удовлетворить эту насущную потребность, и были разработаны различные сложные статистические методы. Несмотря на то что прогнозы, сделанные на основе этих методов, могут оказаться чрезвычайно близкими к оптимальным (если иметь в виду ту ограниченную информацию, которой мы располагаем), они могут не в полной мере удовлетворять ваши реальные потребности.

[1] Исключением является *процесс чистого случайного шума*, описанный в разделе 14.3.

[2] Например, 2 января 1985 г. в *The Wall Street Journal* были опубликованы прогнозы выдающихся экономистов параллельно с реальными результатами развития экономики США. Средний прогноз по процентным ставкам краткосрочных казначейских векселей, составленный на шесть месяцев вперед, равнялся 10,64%. Шесть месяцев спустя оказалось, что фактическая процентная ставка достигла лишь 7,84%. Когда оказывается, что расхождение в процентных ставках, равное одной четверти процента, обходится вашему местному первичному покупателю примерно в $1 000 (в текущих ценах), то подобное расхождение для промышленности и экономики страны в целом выливается в колоссальные суммы. Ни один из этих экономистов, несмотря на все изощренные методы прогнозирования, которыми они пользовались, так и не смог предвидеть подобного поворота событий в изменении процентных ставок.

При анализе временных рядов используется множество различных подходов. Методы анализа временных рядов продолжают совершенствоваться и развиваться. Приведя несколько примеров совокупностей данных, относящихся к категории временных рядов, мы обсудим два из наиболее важных методов анализа временных рядов в сфере бизнеса.

1. *Анализ трендов/сезонности* представляет собой непосредственный, интуитивный подход к оцениванию базовых компонентов помесячных или поквартальных временных рядов. Эти компоненты включают (1) долгосрочную тенденцию (тренд); (2) точное повторение сезонных моделей поведения; (3) среднесрочные ("блуждающие") циклические всплески и падения; (4) случайный, нерегулярный "шум". Прогнозы получаются путем наложения обычных сезонных моделей на долгосрочную тенденцию.
2. *ARIMA-процессы Бокса–Дженкинса* представляют собой гибкие линейные модели, которые могут точно описывать широкий спектр поведения различных временных рядов, в том числе и среднесрочные всплески и падения так называемого "экономического цикла". Несмотря на то что эти базовые модели достаточно просты в описании, их оценка требует обширных компьютерных вычислений. Прогнозы и доверительные границы получаются на основе статистической теории, базирующейся на будущем поведении оцениваемой модели.

Пример. *Фондовая биржа — всего лишь случайное блуждание*

Величина индекса фондовой биржи, регистрируемая ежедневно при закрытии биржи (например, промышленный индекс Доу Джонса), образует временной ряд, который имеет огромное значение для многих из нас. На рис. 14.1.1 представлен типичный график временного ряда, на котором временной ряд откладывается по вертикальной оси Y как функция от времени (количество торговых дней), отложенного по горизонтальной оси X.

Какая информация была бы утрачена, если бы мы начертили гистограмму значений индекса Доу Джонса, вычислили среднее значение или определили обычный доверительный интервал? Мы утратили бы информацию о *последовательности* наблюдений: мы интерпретировали бы значения индекса так, словно последовательность их появления носит совершенно случайный характер. Рис. 14.1.2 показывает, что важная информация утрачивается в том случае, когда предполагается произвольная последовательность. Отсюда можно сделать вывод о необходимости применения в таких случаях специальных методов временных рядов, позволяющих воспользоваться этой важной информацией. Хороший метод временных рядов для данных, касающихся фондовой биржи, должен учитывать то обстоятельство, что показатели фондовой биржи — по мере их обычных колебаний — как правило, меняются день ото дня лишь незначительно (относительно предыдущего дня).

Финансовая теория эффективных рынков утверждает, что поведение фондовой биржи представляет собой *случайное блуждание*, при котором дневные изменения принимают форму непредсказуемого, случайного шума.[3] На рис. 14.1.3 показано, что дневные изменения (сегодняшнее значение минус вчерашнее значение) биржевой цены за этот период времени носят действительно случайный характер.

[3] Поскольку крупные инвесторы, получив интересующую их информацию, действуют немедленно, любые *предсказуемые* тенденции уже отражены в биржевых ценах. Единственно возможными являются те изменения, которые объясняются наличием фактора *непредсказуемости*, или случайности. Более точный анализ должен был бы основываться не на изменениях как таковых, а на процентных изменениях; в таком случае различия присутствуют тогда, когда за рассматриваемый период времени ряд изменился на вполне ощутимый процент от своей исходной величины.

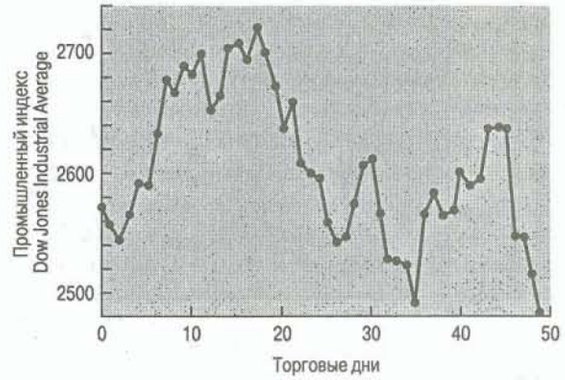


Рис. 14.1.1. График временного ряда для промышленного индекса Доу Джонса, ежедневные значения в период с 31 июля по 9 октября 1987 г.

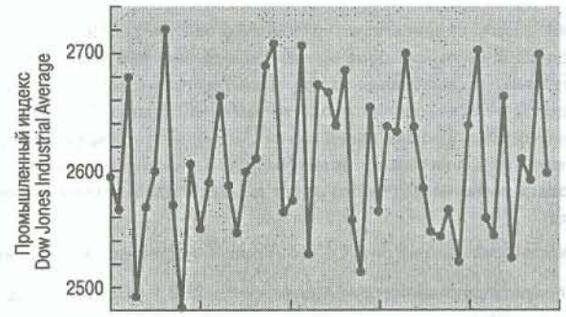


Рис. 14.1.2. Результаты случайного изменения последовательности данных временного ряда из предыдущего рисунка. Утеряна важная информация, поскольку данный график уже не отражает временных тенденций. Анализ временных рядов требует применения специальных методов, которые позволяют сохранить эту важную информацию

Модель *случайных блужданий* включена в ARIMA-структуру Бокса–Дженкинса как особый случай ряда, которому "известно" лишь, в каком состоянии он находится, но "неизвестно", как он оказался в этом состоянии.

Пример. *Производство персональных компьютеров переживает период устойчивого роста*

Компьютерная промышленность переживает период впечатляющего и устойчивого роста. В табл. 14.1.1 представлен рост доходов американских компаний, занимающихся выпуском персональных компьютеров, в период с 1987 по 1995 гг. График временного ряда, показанный на рис. 14.1.4, демонстрирует в целом устойчивый рост производства, причем доходы компьютерной промышленности повышаются каждый год (хотя плавность этого графика несколько нарушена небольшим "всплеском", приходящимся на 1991 г.).

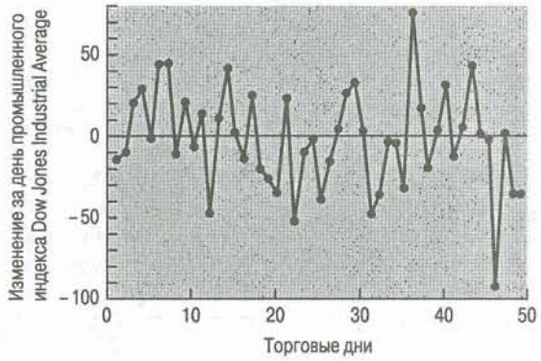


Рис. 14.1.3. Ежесуточные изменения промышленного индекса Доу Джонса, представленные в надлежащей последовательности (без изменения порядка). Это практически случайный ряд, соответствующий модели случайных блужданий для поведения фондовой биржи

Кривизна, направленная вверх, предполагает постоянство темпов роста, что, по сути, означает экспоненциальный рост доходов. Один из способов оценки темпов роста на протяжении этого периода времени заключается в использовании коэффициента регрессии для прогнозирования натурального логарифма данных временного ряда (Y) на основании периода времени (X). В табл. 14.1.2 представлены логарифмы значений этого временного ряда. График логарифмов значений этого временного ряда, показанный на рис. 14.1.5, демонстрирует приблизительно линейную (т.е. в виде прямой линии) зависимость, что лишь подтверждает нашу модель экспоненциального роста доходов компьютерной промышленности (при постоянстве темпов роста).

Оценочная линия регрессии показана на рис. 14.1.6. Уравнение регрессии имеет вид

$$\text{прогнозируемый логарифм доходов} = -282{,}1 + 0{,}1433 \times \text{год}.$$

Каждый дополнительный год добавляет к предыдущему прогнозируемому логарифму значение коэффициента регрессии 0,1433. Таким образом, если воспользоваться экспоненциальной функцией как обратной по

Таблица 14.1.1. Рост доходов американских компаний, занимающихся выпуском персональных компьютеров

Год	Доходы компаний, занимающихся выпуском персональных компьютеров, млрд дол.
1987	14
1988	17
1989	18
1990	19
1991	25
1992	26
1993	30
1994	37
1995	48

Данные взяты из таблицы 1251 и дополнения на компакт-диске *Statistical Abstract of the United States: 1997* (117th edition) Washington, D.C.

отношению к логарифму, новое значение получается путем умножения предыдущего количества на

$$e^{0{,}1433} = 2{,}71828^{0{,}1433} = 1{,}154.$$

Вычитая 1, находим, что оцениваемые темпы роста новых заказов для компьютерной промышленности в период с 1987 по 1995 гг. составляют 15,4% в год.

$$\text{Темпы роста} = 15{,}4\%.$$

Обратите внимание, что точки данных распределены вокруг линии регрессии, показанной на рис. 14.1.6, почти случайно. Говорят, что ряд имеет автокорреляцию, если у точек этого ряда есть тенденция быть несколько выше линии или несколько ниже. Создается впечатление, что в данном примере автокорреляция не представляет особой проблемы. Если же — в иных ситуациях — имеется автокорреляция, то линия наименьших квадратов может по-прежнему служить надежной оценкой роста, однако статистический вывод (доверительные интервалы и проверки гипотез) может привести к неправильным результатам, поскольку автокорреляция недопустима для линейной регрессионной модели, описанной в главе 11.

Таблица 14.1.2. Доходы и логарифмы доходов американских компаний, занимающихся выпуском персональных компьютеров

Год, X	Доходы компаний, занимающихся выпуском персональных компьютеров, млрд дол.	Натуральный логарифм доходов, Y
1987	14	2,64
1988	17	2,83
1989	18	2,89
1990	19	2,94
1991	25	3,22
1992	26	3,26
1993	30	3,40
1994	37	3,61
1995	48	3,87

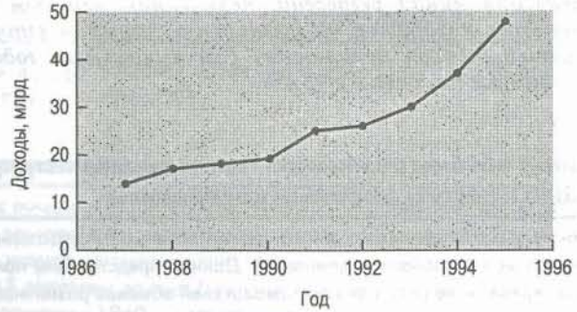


Рис. 14.1.4. Устойчивый рост: доходы американских компаний, занимающихся выпуском персональных компьютеров (с 1987 по 1995 гг.)

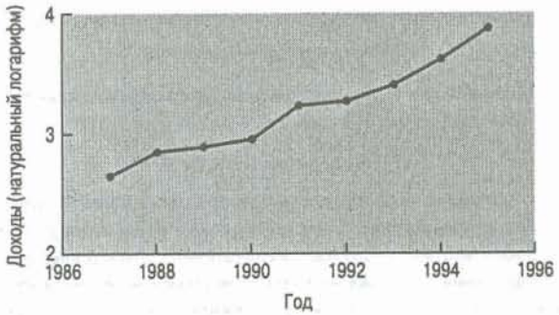


Рис. 14.1.5. Логарифм доходов американских компаний, занимающихся выпуском персональных компьютеров, свидетельствует о линейном (в виде прямой линии) росте с течением времени. Вид графика указывает на то, что темпы роста доходов практически постоянны (т.е. наблюдается экспоненциальный рост)

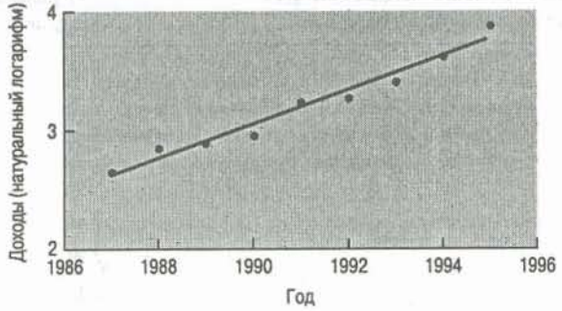


Рис. 14.1.6. Логарифм доходов американских компаний, занимающихся выпуском персональных компьютеров (Y), как функция от времени (X); на рисунке также показана линия регрессии, полученная методом наименьших квадратов. Коэффициент регрессии (наклон, равный 0,1433) использован для вычисления годового темпа роста, равного 15,4%

Пример. *Суммарные объемы розничной торговли демонстрируют тенденцию к росту и сезонным колебаниям*

В табл. 14.1.3 представлены приблизительные и нескорректированные показатели объемов розничной торговли в США (выраженные в миллиардах долларов). Данные представлены помесячно, за период с 1995 по 1997 гг. График временного ряда для этих показателей объемов розничной торговли, представленный на рис. 14.1.7, демонстрирует в целом устойчивый рост с существенными отклонениями при переходе от одного месяца к следующему. Более пристальный анализ этого отклонения показывает, что оно носит отнюдь не случайный характер, а повторяется с довольно высокой точностью в разные годы. Наивысшие точки приходятся на декабрь (как раз перед началом нового года), затем, в январе и феврале, объем продаж падает до минимальных значений. Этот вид сезонных колебаний вполне соответствует нашему пониманию модели совершения покупок накануне рождественских каникул, характерной, в частности, для жителей Соединенных Штатов.

Государственные органы также предоставляют показатели объемов розничной торговли с *поправкой на сезон*, удаляя прогнозируемые изменения между текущим месяцем и следующим за ним, как показано в табл. 14.1.4. Когда из рассматриваемого ряда изымаются предсказуемые сезонные колебания, картина устойчивого роста объемов продаж, как следует из рис. 14.1.8, становится более очевидной (сравните, например, рис. 14.1.8 и 14.1.7). Оставшаяся вариация указывает на флуктуации, не повторяющиеся из года в год. Эти флуктуации отражают изменения, которые явились неожиданностью в соответствующее время года.

Таблица 14.1.3. Объемы розничной торговли в США (без поправки на сезонные колебания)

Год	Месяц	Объем продаж, млрд дол.	Год	Месяц	Объем продаж, млрд дол.
1995	Январь	166	1996	Июль	206
1995	Февраль	163	1996	Август	214
1995	Март	191	1996	Сентябрь	197
1995	Апрель	187	1996	Октябрь	209
1995	Май	200	1996	Ноябрь	212
1995	Июнь	202	1996	Декабрь	246
1995	Июль	194	1997	Январь	188
1995	Август	203	1997	Февраль	185
1995	Сентябрь	192	1997	Март	212
1995	Октябрь	193	1997	Апрель	207
1995	Ноябрь	201	1997	Май	221
1995	Декабрь	237	1997	Июнь	214
1996	Январь	174	1997	Июль	218
1996	Февраль	181	1997	Август	222
1996	Март	201	1997	Сентябрь	209
1996	Апрель	200	1997	Октябрь	218
1996	Май	215	1997	Ноябрь	216
1996	Июнь	206	1997	Декабрь	258

Данные заимствованы в Бюро переписи населения США, `http://www.census.gov/svsd/www/monret.html`, осень 1998 г.



Пример. *Процентные ставки*

Одним из способов, с помощью которых правительство США получает деньги, является продажа ценных бумаг. Казначейские векселя представляют собой краткосрочные ценные бумаги со сроком погашения не более одного года (в момент погашения владельцу выплачивается их полная стоимость плюс процентная ставка). В табл. 14.1.5 показаны доходы (процентные ставки) по трехмесячным казначейским векселям США за каждый год, начиная с 1960 г. и заканчивая 1997 г. включительно.

График соответствующего временного ряда, показанный на рис. 14.1.9, демонстрирует общий подъем и последующее снижение процентных ставок за рассматриваемый период времени со значительной вариацией. Заметна определенная цикличность подъемов и снижений процентных ставок с нарастанием величины. Пытаясь интерпретировать эту модель поведения процентных ставок, следует, однако, проявлять осторожность. Учитывая вмешательство правительства, экономисты в целом не рассчитывают на то, что подобная модель роста процентных ставок сохранится и в будущем.

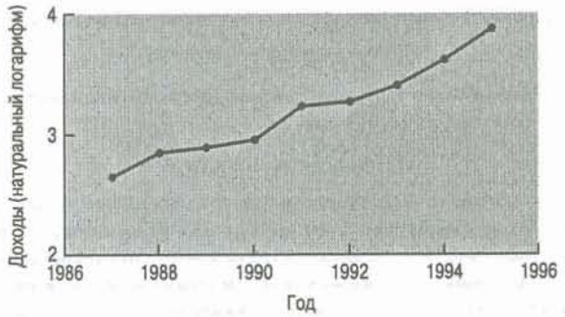


Рис. 14.1.5. Логарифм доходов американских компаний, занимающихся выпуском персональных компьютеров, свидетельствует о линейном (в виде прямой линии) росте с течением времени. Вид графика указывает на то, что темпы роста доходов практически постоянны (т.е. наблюдается экспоненциальный рост)

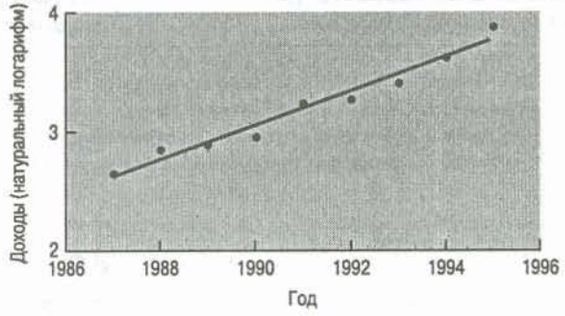


Рис. 14.1.6. Логарифм доходов американских компаний, занимающихся выпуском персональных компьютеров (Y), как функция от времени (X); на рисунке также показана линия регрессии, полученная методом наименьших квадратов. Коэффициент регрессии (наклон, равный 0,1433) использован для вычисления годового темпа роста, равного 15,4%

Пример. *Суммарные объемы розничной торговли демонстрируют тенденцию к росту и сезонным колебаниям*

В табл. 14.1.3 представлены приблизительные и нескорректированные показатели объемов розничной торговли в США (выраженные в миллиардах долларов). Данные представлены помесячно, за период с 1995 по 1997 гг. График временного ряда для этих показателей объемов розничной торговли, представленный на рис. 14.1.7, демонстрирует в целом устойчивый рост с существенными отклонениями при переходе от одного месяца к следующему. Более пристальный анализ этого отклонения показывает, что оно носит отнюдь не случайный характер, а повторяется с довольно высокой точностью в разные годы. Наивысшие точки приходятся на декабрь (как раз перед началом нового года), затем, в январе и феврале, объем продаж падает до минимальных значений. Этот вид сезонных колебаний вполне соответствует нашему пониманию модели совершения покупок накануне рождественских каникул, характерной, в частности, для жителей Соединенных Штатов.

Государственные органы также предоставляют показатели объемов розничной торговли *с поправкой на сезон*, удаляя прогнозируемые изменения между текущим месяцем и следующим за ним, как показано в табл. 14.1.4. Когда из рассматриваемого ряда изымаются предсказуемые сезонные колебания, картина устойчивого роста объемов продаж, как следует из рис. 14.1.8, становится более очевидной (сравните, например, рис. 14.1.8 и 14.1.7). Оставшаяся вариация указывает на флуктуации, не повторяющиеся из года в год. Эти флуктуации отражают изменения, которые явились неожиданностью в соответствующее время года.

Таблица 14.1.3. Объемы розничной торговли в США (без поправки на сезонные колебания)

Год	Месяц	Объем продаж, млрд дол.	Год	Месяц	Объем продаж, млрд дол.
1995	Январь	166	1996	Июль	206
1995	Февраль	163	1996	Август	214
1995	Март	191	1996	Сентябрь	197
1995	Апрель	187	1996	Октябрь	209
1995	Май	200	1996	Ноябрь	212
1995	Июнь	202	1996	Декабрь	246
1995	Июль	194	1997	Январь	188
1995	Август	203	1997	Февраль	185
1995	Сентябрь	192	1997	Март	212
1995	Октябрь	193	1997	Апрель	207
1995	Ноябрь	201	1997	Май	221
1995	Декабрь	237	1997	Июнь	214
1996	Январь	174	1997	Июль	218
1996	Февраль	181	1997	Август	222
1996	Март	201	1997	Сентябрь	209
1996	Апрель	200	1997	Октябрь	218
1996	Май	215	1997	Ноябрь	216
1996	Июнь	206	1997	Декабрь	258

Данные заимствованы в Бюро переписи населения США, `http://www.census.gov/svsd/www/monret.html`, осень 1998 г.



Пример. *Процентные ставки*

Одним из способов, с помощью которых правительство США получает деньги, является продажа ценных бумаг. Казначейские векселя представляют собой краткосрочные ценные бумаги со сроком погашения не более одного года (в момент погашения владельцу выплачивается их полная стоимость плюс процентная ставка). В табл. 14.1.5 показаны доходы (процентные ставки) по трехмесячным казначейским векселям США за каждый год, начиная с 1960 г. и заканчивая 1997 г. включительно.

График соответствующего временного ряда, показанный на рис. 14.1.9, демонстрирует общий подъем и последующее снижение процентных ставок за рассматриваемый период времени со значительной вариацией. Заметна определенная цикличность подъемов и снижений процентных ставок с нарастанием величины. Пытаясь интерпретировать эту модель поведения процентных ставок, следует, однако, проявлять осторожность. Учитывая вмешательство правительства, экономисты в целом не рассчитывают на то, что подобная модель роста процентных ставок сохранится и в будущем.

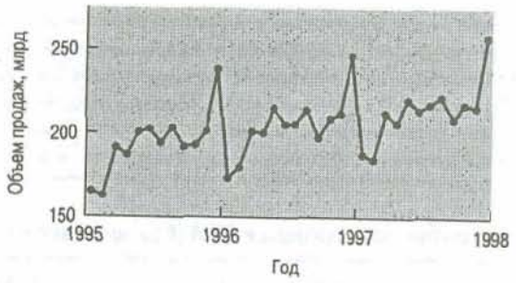


Рис. 14.1.7. Помесячные объемы розничной торговли в США с 1995 по 1997 гг. включительно. Обратите внимание на значительные сезонные колебания, повторяющиеся из года в год, а также на устойчивый рост объемов розничной торговли в целом

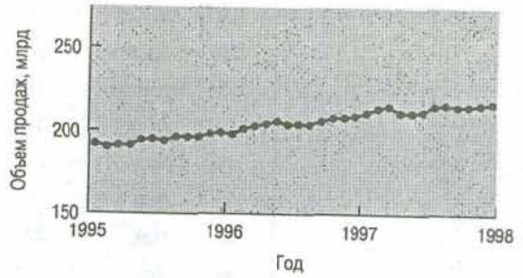


Рис. 14.1.8. Помесячные объемы розничной торговли в США с поправкой на сезонные колебания с 1995 по 1997 гг. включительно. Изъяв из графика сезонные колебания, нам удалось сделать более очевидным устойчивый рост объемов розничной торговли в целом. Оставшаяся вариация указывает на изменения, которые явились неожиданностью в соответствующее время года

Таблица 14.1.4. Объемы розничной торговли в США (с поправкой на сезонные колебания)

		Объем продаж, млрд дол.				Объем продаж, млрд дол.	
Год	Месяц	Без поправки	С поправкой на сезонные колебания	Год	Месяц	Без поправки	С поправкой на сезонные колебания
1995	Январь	166	192	1996	Июль	206	204
1995	Февраль	163	190	1996	Август	214	204
1995	Март	191	191	1996	Сентябрь	197	206
1995	Апрель	187	191	1996	Октябрь	209	208
1995	Май	200	194	1996	Ноябрь	212	208
1995	Июнь	202	195	1996	Декабрь	246	209

		Объем продаж, млрд дол.				Объем продаж, млрд дол.	
Год	Месяц	Без поправки	С поправкой на сезонные колебания	Год	Месяц	Без поправки	С поправкой на сезонные колебания
1995	Июль	194	194	1997	Январь	188	211
1995	Август	203	196	1997	Февраль	185	214
1995	Сентябрь	192	196	1997	Март	212	215
1995	Октябрь	193	196	1997	Апрель	207	211
1995	Ноябрь	201	198	1997	Май	221	211
1995	Декабрь	237	199	1997	Июнь	214	212
1996	Январь	174	198	1997	Июль	218	215
1996	Февраль	181	201	1997	Август	222	216
1996	Март	201	203	1997	Сентябрь	209	215
1996	Апрель	200	204	1997	Октябрь	218	215
1996	Май	215	206	1997	Ноябрь	216	216
1996	Июнь	206	204	1997	Декабрь	258	217

Данные заимствованы в Бюро переписи населения США, `http://www.census.gov/svsd/www/monret.html`, осень 1998 г.



Таблица 14.1.5. Казначейские векселя США (трехмесячный срок погашения)

Год	Доход	Год	Доход
1960	2,93	1979	10,05
1961	2,38	1980	11,51
1962	2,78	1981	14,03
1963	3,16	1982	10,69
1964	3,56	1983	8,63
1965	3,95	1984	9,35
1966	4,88	1985	7,47
1967	4,32	1986	5,98
1968	5,34	1987	5,82
1969	6,68	1988	6,69
1970	6,43	1989	8,12
1971	4,35	1990	7,51
1972	4,07	1991	5,42
1973	7,04	1992	3,45
1974	7,89	1993	3,02

Окончание табл. 14.1.5

Год	Доход	Год	Доход
1975	5,84	1994	4,29
1976	4,99	1995	5,51
1977	5,27	1996	5,02
1978	7,22	1997	5,07

Данные заимствованы в Совете управляющих Федеральной резервной системы США, http://www.bog.frb.fed.us/releases/H15/data/a/tbaa3m.txt, осень 1998 г.



В отличие от примера с доходами предприятий компьютерной промышленности, трудно рассчитывать, что процентные ставки и в дальнейшем будут непрерывно расти. В отличие от примера с объемами розничной торговли, циклы процентных ставок на рис. 14.1.9 не демонстрируют строгого повторения.

Временной ряд блуждающего характера зачастую порождает тренды (тенденции) и циклы, которые вовсе не обязательно сохранятся и в будущем. Подход на основе ARIMA-процесса Бокса–Дженкинса (который мы обсудим более подробно в разделе 14.3) особенно хорошо приспособлен именно для такого типа поведения временных рядов, поскольку он учитывает то обстоятельство, что ряд обычно порождает циклы в случае, когда он имеет блуждающий характер.

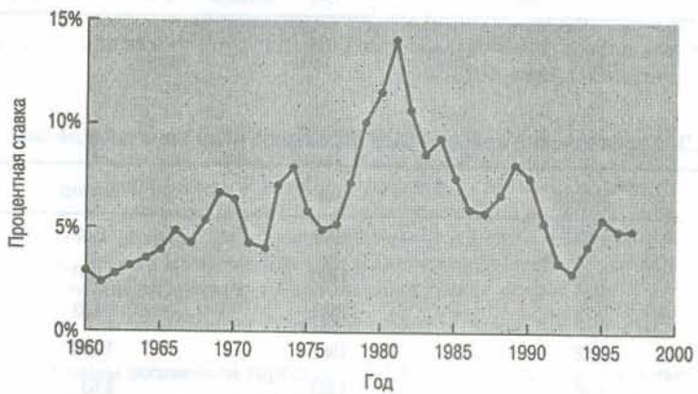


Рис. 14.1.9. Процентные ставки по трехмесячным казначейским векселям США с 1960 по 1997 гг. включительно в целом демонстрируют тенденцию к нарастанию (с последующим убыванием) и циклическим флуктуациям. Эти циклы, однако, имеют разную продолжительность; к тому же трудно рассчитывать, что они будут повторяться в точности так же в дальнейшем, как это имеет место в случае сезонных колебаний

14.2. Анализ трендов и сезонности

Анализ трендов и сезонности представляет собой непосредственный, интуитивный подход к оцениванию четырех базовых компонентов помесячных или поквартальных временных рядов: долгосрочный тренд (тенденция), сезонность, циклическая вариация и нерегулярный компонент. Базовая модель временного ряда представляет числа в этом ряде в виде произведения, получаемого путем умножения перечисленных компонентов.

Модель временного ряда, основанная на трендах и сезонности

Данные = тренд × сезонность × цикличность × нерегулярность.

Ниже приведены определения этих четырех базовых компонентов.

1. Долгосрочный **тренд (тенденция)** указывает *действительно* долгосрочное поведение временного ряда — как правило, в виде прямой линии или экспоненциальной кривой. Это бывает полезно в случае, если требуется увидеть картину в целом.
2. Точно повторяющийся **сезонный компонент** определяет влияние времени года. Например, потребность в устройствах обогрева высока в зимние месяцы, соответствующие объемы продаж высоки в декабре, а объемы продаж сельскохозяйственной продукции высоки в период сбора урожая. Каждый период времени в течение года характеризуется своим *сезонным индексом*, который свидетельствует о том, насколько выше или ниже соответствующий показатель в данный период времени по сравнению с другими периодами. Например, в случае поквартальных данных имеется сезонный индекс для каждого квартала: индекс, равный 1,235, для четвертого квартала, говорит о том, что объем продаж в четвертом квартале примерно на 23,5% выше, чем в первом, втором или третьем кварталах. Индекс, равный 0,921, для второго квартала, говорит о том, что объем продаж во втором квартале ниже примерно на 7,9% (поскольку 1 – 0,921 = 0,079).
3. Среднесрочный **циклический компонент** состоит из последовательных повышений и понижений, которые *не* повторяются каждый год и поэтому исключаются из сезонного компонента. Поскольку эти повышения и понижения чередуются, их нельзя считать достаточно случайными и рассматривать как часть независимой случайной ошибки (нерегулярного компонента). Циклическую вариацию особенно трудно прогнозировать за пределами ближайшего будущего. Тем не менее она может быть очень важна, поскольку основные явления экономического цикла (такие как экономический спад) рассматриваются как часть циклической вариации в экономических показателях.
4. Краткосрочный **нерегулярный (случайный) компонент** представляет остаточную вариацию, которую невозможно объяснить. В нем проявляется дей-

ствие тех однократных событий, которые происходят с течением времени случайно, а не систематически. Самое большее, что можно сделать с этим нерегулярным компонентом, оценить его величину (воспользовавшись, например, стандартным отклонением), определить, меняется ли он с течением времени, и признать, что даже в идеальных условиях прогноз не может быть точнее (в среднем), чем типичная величина нерегулярной вариации.

Эти четыре базовых компонента временного ряда (тренд, сезонность, циклический и нерегулярный компоненты) можно оценивать различными способами. Ниже приведен краткий обзор метода, который называется **отношением к скользящему среднему** (позже мы обсудим его более подробно). В основе этого метода лежит деление значений ряда на гладкое скользящее среднее следующим образом.

1. *Скользящее среднее* используется для устранения сезонных эффектов путем усреднения по всему году, для уменьшения нерегулярного компонента и получения комбинации тренда и циклического компонента.
2. Деление исходного ряда на сглаженный ряд скользящего среднего дает нам *отношение к скользящему среднему*, которое включает как сезонные, так и нерегулярные значения. Выполняя группирование по времени года, а затем усреднение в полученных группах, находим *сезонный индекс* для каждого времени года. Выполняя деление каждого значения ряда на соответствующий сезонный индекс для соответствующего времени года, находим значения *с сезонной поправкой*.
3. Регрессия ряда с сезонной поправкой (Y) по времени (X) служит для оценивания *долгосрочного тренда* в виде прямой линии как функции от времени.[4] Этот тренд (тенденция) не отражает сезонных колебаний и дает возможность получить прогноз с сезонной поправкой.
4. Прогнозирование можно выполнять с помощью *сезонности тренда*. Получая из уравнения регрессии прогнозируемые значения (тренд) для будущих периодов времени и затем умножая их на соответствующий сезонный индекс, вы получаете прогнозы, которые отражают как долгосрочную тенденцию, так и сезонное поведение.

Преимуществом метода отношения к скользящему среднему является легкость вычислений и интерпретации результатов. Основной недостаток заключается в том, что соответствующая модель не является полностью определенной; как следствие, бывает довольно сложно найти меры неопределенности (например, границы прогноза).[5]

Приведенный ниже пример временного ряда позволяет более подробно ознакомиться со всеми этими компонентами. Далее в этом разделе мы не раз будем возвращаться к этому примеру.

[4] Например, эта переменная времени X может состоять из чисел 1, 2, 3,... .

[5] На практике подробности такой частично случайной структуры циклического компонента не указываются. Эту проблему не решить с помощью метода множественной регрессии, в котором для оценивания сезонных индексов используются индикаторные переменные.

Пример. *Продажа автомобилей Ford Motor Company*

В табл. 14.2.1 представлены поквартальные объемы продаж автомобилей Ford Motor Company (по материалам этой компании). Этот временной ряд демонстрирует ярко выраженные сезонные колебания. Объемы продажи, как правило, достигают пика во втором квартале, о чем свидетельствует график временного ряда для соответствующих данных (рис. 14.2.1). Затем они в целом нарастают в течение последующих трех кварталов.[6] Поскольку этот сезонный сценарий не повторяется в точности каждый год, рассматриваемый временной ряд характеризуется также некоторой цикличностью и нерегулярностью поведения. Советуем обратить внимание и на долгосрочную тенденцию, выражающуюся в общем росте продаж с течением времени (за исключением, возможно, конца временного ряда).

Результаты анализа тренда и сезонности представлены на рис. 14.2.2. Соответствующий тренд выражается прямой линией, сезонный индекс повторяется в точности каждый год, циклический компонент непостоянен, а нерегулярный компонент носит в основном случайный характер. Поскольку циклический и нерегулярный компоненты относительно невелики по сравнению с сезонными колебаниями, их масштаб на рис. 14.2.3 несколько увеличен, чтобы показать их цикличность и нерегулярность. Ниже мы дадим пояснения к вычислениям для этого анализа; сейчас же вы должны понять, как эти базовые компоненты соотносятся с исходным временным рядом.

Тренд и циклический компонент: скользящее среднее

Наша цель заключается в том, чтобы выделить четыре базовых компонента временного ряда. Начнем с усреднения данных за год, чтобы избавиться от сезонного компонента и уменьшить нерегулярный компонент. **Скользящее среднее** представляет собой новый ряд, полученный путем усреднения соседних наблюдений временного ряда и перехода к следующему периоду времени — в итоге получается более гладкий ряд. Выполняя усреднение данных за целый год, мы приходим к тому, что вклад сезонных компонентов — независимо от времени года — остается практически одинаковым.

Скользящее среднее = тренд × цикличность.

Таблица 14.2.1. Ford Motor Company

Год	Квартал	Объем продаж автомобилей, млн дол.	Год	Квартал	Объем продаж автомобилей, млн дол.
1991	1	17 115	1994	3	24 926
1991	2	19 833	1994	4	27 766
1991	3	17 205	1995	1	28 601
1991	4	17 898	1995	2	29 861
1992	1	20 636	1995	3	24 437
1992	2	22 903	1995	4	27 597
1992	3	19 370	1996	1	28 297
1992	4	21 498	1996	2	31 505

[6] Единственным исключением из этого правила является падение объемов продаж, наблюдавшееся с четвертого квартала 1996 г. по первый квартал 1997 г.

Окончание табл. 14.2.1

Год	Квартал	Объем продаж автомобилей, млн дол.	Год	Квартал	Объем продаж автомобилей, млн дол.
1993	1	22 686	1996	3	26 459
1993	2	25 264	1996	4	31 505
1993	3	20 107	1997	1	30 037
1993	4	23 511	1997	2	32 805
1994	1	26 070	1997	3	28 196
1994	2	28 375	1997	4	31 897

Данные получены из ежегодных отчетов компании, Детройт, шт. Мичиган.

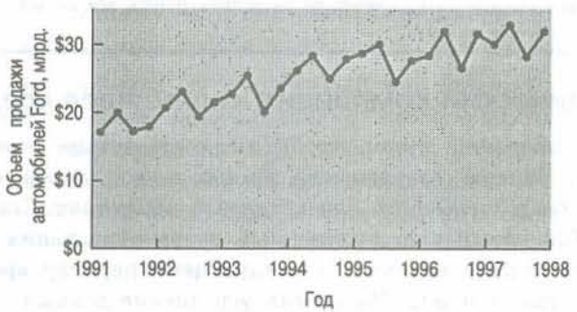


Рис. 14.2.1. График временного ряда поквартальных продаж автомобилей компании Ford Motor Company. Обратите внимание на повторяющиеся ежегодно сильные сезонные колебания. Можно также заметить долговременную тенденцию к общему повышению объемов продажи на протяжении большей части рассматриваемого отрезка времени, а также некоторую нерегулярность поведения

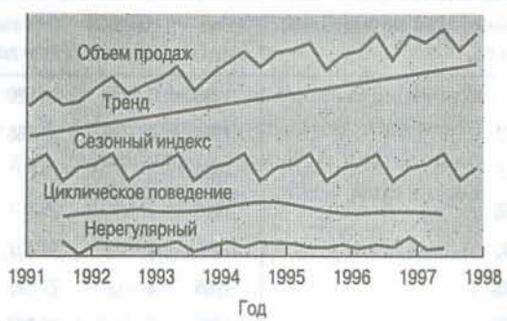


Рис. 14.2.2. График поквартальных продаж с разбивкой на четыре базовых компонента: тренд (представлен прямой линией), сезонный индекс, который повторяется каждый год, "блуждающий" циклический компонент и случайный нерегулярный компонент. Все эти компоненты показаны в том же масштабе, что и исходный ряд



Рис. 14.2.3. Циклический и нерегулярный компоненты представлены в увеличенном масштабе, чтобы продемонстрировать подробности их поведения

Найти скользящее среднее значение для поквартальных данных за определенный период времени можно следующим образом. Начните с текущего значения, добавьте к нему значения его "соседей", затем добавьте *половину* значений следующих "соседей" и разделите на 4. Такое взвешенное среднее необходимо для того, чтобы интервал по обе стороны от базового периода времени был симметричным и вместе с тем охватывал в точности данные за один год.[7] Если вы располагаете помесячными данными, усредните соответствующее значение в базовый период времени вместе с пятью ближайшими месяцами с каждой стороны и половиной следующего по счету месяца (вслед за пятью ближайшими) — также с каждой стороны. Скользящее среднее отсутствует для первых двух и последних двух кварталов ряда или, если речь идет о помесячных данных, для первых шести и последних шести месяцев ряда.

В случае Ford Motor Company скользящее среднее объема продаж автомобилей за третий квартал 1991 г. вычисляется следующим образом: $[(1/2) \times 17\,115 + 19\,833 + 17\,205 + 17\,898 + (1/2) \times 20\,636]/4 = 18\,453$. За четвертый квартал 1991 г. скользящее среднее объема продаж составляет $[(1/2) \times 19\,833 + 17\,205 + 17\,898 + 20\,636 + (1/2) \times 22\,903]/4 = 19\,277$. Значения скользящего среднего показаны в табл. 14.2.2, а в графическом виде представлены на графике временного ряда (рис. 14.2.4).

Сезонный индекс: среднее значение отношения к скользящему среднему отражает сезонное поведение

Чтобы выделить сезонное поведение, прежде всего следует получить отношение исходных значений к скользящему среднему. (Именно отсюда происходит название "отношение к скользящему среднему".) Полученный результат будет включать сезонный и нерегулярный компоненты, поскольку скользящее среднее исключает из данных тренд и циклический компонент.

$$\text{(Сезонность) (Нерегулярность)} = \frac{\text{Данные}}{\text{Скользящее среднее}}$$

[7] Взвешивая крайние точки коэффициентом 1/2, вы гарантируете, что этот квартал учтен в скользящем среднем точно так же, как и другие кварталы.

Таблица 14.2.2. Объемы продаж автомобилей компании Ford Motor Company со скользящим средним

Год	Квартал	Объем продаж автомобилей, млн дол.	Скользящее среднее объема продаж автомобилей, млн дол.
1991	1	17 115	(отсутствует)
1991	2	19 833	(отсутствует)
1991	3	17 205	18 453
1991	4	17 898	19 277
1992	1	20 636	19 931
1992	2	22 903	20 652
1992	3	19 370	21 358
1992	4	21 498	21 909
1993	1	22 686	22 297
1993	2	25 264	22 640
1993	3	20 107	23 315
1993	4	23 511	24 127
1994	1	26 070	25 118
1994	2	28 375	26 252
1994	3	24 926	27 101
1994	4	27 766	27 603
1995	1	28 601	27 727
1995	2	29 861	27 645
1995	3	24 437	27 586
1995	4	27 597	27 786
1996	1	28 297	28 276
1996	2	31 762	29 017
1996	3	26 459	29 723
1996	4	31 505	30 071
1997	1	30 037	30 419
1997	2	32 805	30 685
1997	3	28 196	(отсутствует)
1997	4	31 897	(отсутствует)

Затем, чтобы устранить нерегулярный компонент, вы усредняете эти значения для каждого сезона. Сезонный компонент проявляется, поскольку он присутствует ежегодно, тогда как нерегулярный компонент, как правило, удается усреднить. Конечные результаты включают **сезонный индекс** для каждого времени года — фактор, который указывает, насколько большим (или меньшим) бывает рассматриваемый показатель в этот конкретный период времени в сравнении с типичным периодом на протяжении года. Например, сезонный индекс за первый квартал, равный 1,30, свидетельствует о том, что рассматриваемый показатель в первом квартале, как правило, на 30% больше, чем в типичном квартале. С другой стороны, сезонный индекс за третий квартал, равный 0,74, свидетельствует о том, что рассматриваемый показатель в третьем квартале, как правило, на 26% ниже, чем в типичном квартале.

$$\text{Сезонный индекс} = \text{Среднее значение}\left(\frac{\text{Данные}}{\text{Скользящееe среднее}}\right)\text{за соответствующий сезон}$$

Для данных Ford Motor Company первое значение отношения к скользящему среднему за третий квартал 1991г. составляет: 17 205/18 453 = 0,93237. Сезонный индекс за третий квартал определяется путем усреднения значений этих отношений за третий квартал по всем рассматриваемым годам, как показано в табл. 14.2.3.

После того как вычислен каждый сезонный индекс, его можно использовать везде — даже там, где нельзя вычислить скользящее среднее, поскольку, по определению, сезонные колебания в точности повторяются каждый год. В табл. 14.2.4 представлены значения отношения к скользящему среднему и сезонные индексы для данных Ford Motor Company. Типичная картина за год показана на рис. 14.2.5, а повторяющиеся сезонные колебания — на рис. 14.2.6.

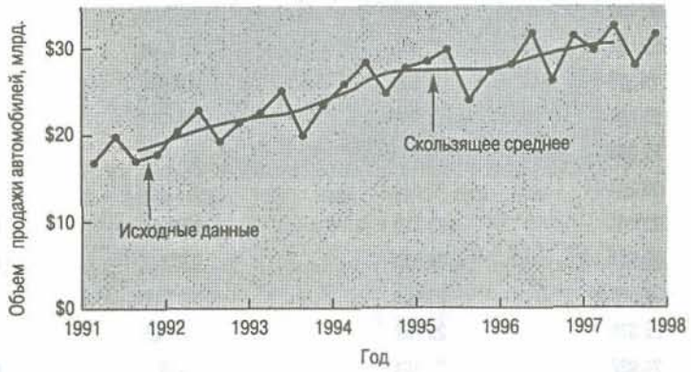


Рис. 14.2.4. Скользящее среднее объемов продажи автомобилей компании Ford Motor Company. Из графика удалось устранить сезонные и нерегулярные колебания объемов продажи; остались лишь тренд и циклический компонент

Таблица 14.2.3. Вычисление сезонного индекса за третий квартал для Ford Motor Company

Год	Отношение к скользящему среднему за третий квартал
1991	0,93237
1992	0,90692
1993	0,86241
1994	0,91976
1995	0,88585

Окончание табл. 14.2.3

Год	Отношение к скользящему среднему за третий квартал
1996	0,89018
1997	(отсутствует)
Среднее	0,8996

Таблица 14.2.4. Объемы продаж автомобилей компании Ford Motor Company и сезонные индексы

Год	Квартал	Объем продаж автомобилей, млн дол.	Скользящее среднее объема продаж автомобилей, млн дол.	Отношение к скользящему среднему	Сезонный индекс
1991	1	17 115	(отсутствует)	(отсутствует)	1,0184
1991	2	19 833	(отсутствует)	(отсутствует)	1,0916
1991	3	17 205	18 453	0,93237	0,8996
1991	4	17 898	19 277	0,92848	0,9885
1992	1	20 636	19 931	1,03537	1,0184
1992	2	22 903	20 652	1,10901	1,0916
1992	3	19 370	21 358	0,90692	0,8996
1992	4	21 498	21 909	0,98122	0,9885
1993	1	22 686	22 297	1,01746	1,0184
1993	2	25 264	22 640	1,11588	1,0916
1993	3	20 107	23 315	0,86241	0,8996
1993	4	23 511	24 127	0,97447	0,9885
1994	1	26 070	25 118	1,03790	1,0184
1994	2	28 375	26 252	1,08085	1,0916
1994	3	24 926	27 101	0,91976	0,8996
1994	4	27 766	27 603	1,00591	0,9885
1995	1	28 601	27 727	1,03151	1,0184
1995	2	29 861	27 645	1,08015	1,0916
1995	3	24 437	27 586	0,88585	0,8996
1995	4	27 597	27 786	0,99321	0,9885
1996	1	28 297	28 276	1,00074	1,0184
1996	2	31 762	29 017	1,09459	1,0916
1996	3	26 459	29 723	0,89018	0,8996
1996	4	31 505	30 071	1,04768	0,9885
1997	1	30 037	30 419	0,98745	1,0184
1997	2	32 805	30 685	1,06910	1,0916
1997	3	28 196	(отсутствует)	(отсутствует)	0,8996
1997	4	31 897	(отсутствует)	(отсутствует)	0,9885

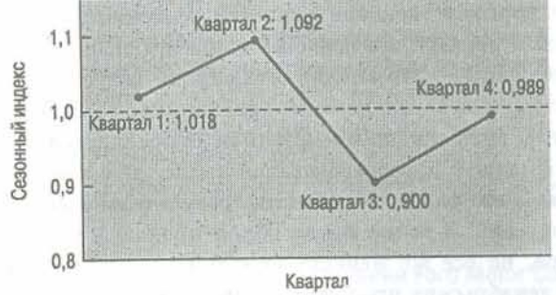


Рис. 14.2.5. Сезонные индексы показывают, что объемы продаж автомобилей компании Ford Motor Company, как правило, достигают пика во втором квартале, падают до минимума в третьем квартале, а затем снова повышаются вплоть до следующего второго квартала

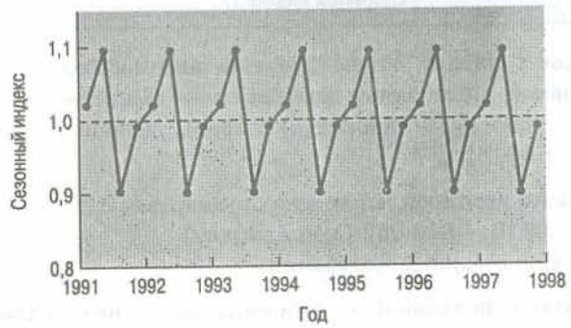


Рис. 14.2.6. Сезонный компонент объемов продажи автомобилей компании Ford Motor Company, извлеченный из исходного ряда, в точности повторяется каждый год

Поправка на сезон: деление ряда на сезонный индекс

3 июня 1998 г. на первой странице *The Wall Street Journal* были опубликованы следующие статистические данные с поправкой на сезонные колебания.

"По сведениям Министерства экономики, объем продажи новых домов, рассчитанных на одну семью, повысился (с поправкой на сезонные колебания) с годового уровня 844 000 в марте до 888 000 в апреле".

Что означают слова "с поправкой на сезонные колебания" и может ли наблюдаться снижение показателя, вычисленного с поправкой на сезонные колебания, даже в том случае, когда его фактическое значение повышается? **Поправка на сезонные колебания** устраняет из результатов измерения ожидаемый сезонный компонент (путем деления ряда на сезонный индекс для соответствующего периода), что позволяет нам непосредственно сравнивать один квартал или месяц с другим (после внесения поправки на сезон), выявляя таким образом те или иные скрытые тенденции.

Для розничной торговли декабрь является особенно благоприятным месяцем. Если объем продаж в декабре оказывается выше по сравнению с ноябрем, в этом нет ничего удивительного — это вполне ожидаемый результат. Но если объем продаж в декабре оказывается выше даже по сравнению с ожидаемыми показателями, самое время распить бутылку шампанского и отправляться в теплые края. Сказать, что "объем продаж в декабре оказался выше, чем в ноябре, с поправкой на сезонные колебания", все равно что сказать "результаты в декабре оказались даже выше, чем мы ожидали". В то же время объем продаж в декабре может оказаться выше, чем в ноябре, но все же меньше ожидаемого, а значит, с поправкой на сезонные колебания декабрьские продажи на самом деле *снижаются*.

Чтобы найти некоторое значение с поправкой на сезонные колебания, достаточно разделить исходные данные на сезонный индекс для соответствующего месяца или квартала.

$$\text{Значение с поправкой на сезон} = \left(\frac{\text{Данные}}{\text{Сезонный индекс}}\right) = \text{Тренд} \times \text{Цикличность} \times \text{Нерегулярность}$$

Для Ford Motor Company объем продажи автомобилей во втором квартале 1997г. — с поправкой на сезонные колебания — вычисляется как фактический объем продажи (32 805, в миллионах долларов), деленный на сезонный индекс второго квартала (1,0916).

Объем продажи автомобилей во втором квартале 1997 г.=
= \$32 805/1,0916 = \$30 052 (в миллионах)
(с поправкой на сезонные колебания).

Почему результат с поправкой на сезонные колебания оказался меньше фактического объема продаж? Дело в том, что объем продажи во втором квартале, как правило, выше по сравнению с типичным кварталом года. В сущности, вы заранее рассчитываете на то, что объем продажи во втором квартале будет примерно на 9,2% выше (исходя из сезонного индекса, равного 1,0916). Деление на сезонный индекс нивелирует влияние этой ожидаемой сезонной флуктуации, приводя объем продажи во втором квартале в соответствие с типичным кварталом года (т.е. снижая его).

В следующем квартале (третий квартал 1997г.) объем продажи с поправкой на сезонные колебания равняется 28 196/0,8996 = 31 343. Обратите внимание на резкое падение объема продаж (с 32 805 во втором квартале до 28 196 в третьем квартале 1997 г.). Однако если воспользоваться поправкой на сезонные колебания, то оказывается, что объем продажи даже повысился, — с 30 052 до 31 343. Это свидетельствует о том, что отмеченное нами серьезное (на первый взгляд) падение объема продаж на самом деле оказалось меньше, чем можно было бы ожидать для этого времени года.

Обратите внимание на значительное увеличение объема продаж в четвертом квартале 1997 г. (с 28 196 до 31 897). Если воспользоваться поправкой на сезонные колебания, то оказывается, что и в этом случае наблюдается рост (с 31 343 до 32 268). Сезонная поправка лишь подтверждает нашу догадку о том, что в этом случае мы имеем дело с "настоящим" ростом объема продаж, а не просто с их сезонным увеличением.

В табл. 14.2.5 представлены объемы продажи с поправкой на сезонные колебания для всего временного ряда. В графическом виде эти данные представлены на рис. 14.2.7 (вместе с исходными данными). Ряд, в котором учитывается поправка на сезонные колебания, оказывается несколько более гладким, чем исходные данные, поскольку в первом случае нам удалось избавиться от сезонных флуктуаций. Однако и в этом случае остаются немалые "шероховатости", поскольку, помимо тренда, в нем по-прежнему присутствуют нерегулярный и циклический компоненты.

Таблица 14.2.5. Объемы продаж автомобилей компании Ford Motor Company и объемы продаж с поправкой на сезонные колебания

Год	Квартал	Объем продаж автомобилей, млн дол.	Сезонный индекс	Объем продаж с поправкой на сезонные колебания, млн дол.
1991	1	17 115	1,0184	16 806
1991	2	19 833	1,0916	18 169
1991	3	17 205	0,8996	19 126
1991	4	17 898	0,9885	18 106
1992	1	20 636	1,0184	20 263
1992	2	22 903	1,0916	20 981
1992	3	19 370	0,8996	21 532
1992	4	21 498	0,9885	21 748
1993	1	22 686	1,0184	22 276
1993	2	25 264	1,0916	23 144
1993	3	20 107	0,8996	22 352
1993	4	23 511	0,9885	23 785
1994	1	26 070	1,0184	25 599
1994	2	28 375	1,0916	25 994
1994	3	24 926	0,8996	27 708
1994	4	27 766	0,9885	28 089
1995	1	28 601	1,0184	28 084
1995	2	29 861	1,0916	27 355
1995	3	24 437	0,8996	27 165
1995	4	27 597	0,9885	27 918
1996	1	28 297	1,0184	27 786
1996	2	31 762	1,0916	29 097
1996	3	26 459	0,8996	29 413
1996	4	31 505	0,9885	31 872
1997	1	30 037	1,0184	29 494
1997	2	32 805	1,0916	30 052
1997	3	28 196	0,8996	31 343
1997	4	31 897	0,9885	32 268



Рис. 14.2.7. Ряд, отражающий объемы продаж с поправкой на сезонные колебания, позволяет вам сравнивать один квартал с другим. Избавляясь от ожидаемых сезонных флуктуаций, мы получаем более четкую картину развития своего бизнеса

Долгосрочный тренд и прогноз с поправкой на сезонные колебания: линия регрессии

Когда временной ряд демонстрирует долгосрочную линейную тенденцию к нарастанию или снижению, для оценки этой тенденции и прогнозирования будущего можно воспользоваться регрессионным анализом. Несмотря на то что такой подход дает возможность получать полезные прогнозы, еще более точный и сложный метод (например, *ARIMA-процесс*) позволяет уделить циклическому компоненту больше внимания, чем представленный здесь метод.

Регрессионный анализ в этом случае сводится к следующему. Для прогнозирования ряда, в котором учитывается поправка на сезонные колебания (переменная Y), используется период времени (переменная X).[8] Результирующее уравнение регрессии будет представлять долгосрочный тренд. Подставляя будущие временные периоды в качестве новых значений X, вы получите возможность экстраполировать эту долгосрочную тенденцию на будущее.

Способу представления временных периодов следует уделить особое внимание. Очень важно, чтобы выбранные вами числа были равномерно распределены.[9]

[8] Если ваш ряд демонстрирует не линейную взаимосвязь, а значительный экспоненциальный рост (что нередко наблюдается во вновь созданных фирмах), то в качестве переменной Y можно воспользоваться *логарифмом* ряда, учитывающего поправку на сезонные колебания, а затем выполнить обратное преобразование предсказанных значений (см. главу 12) для получения прогнозов.

[9] Вы наверняка *не* захотите использовать числа 1991,1; 1991,2; 1991,3; 1991,4; 1992,1;..., поскольку они распределены неравномерно. Вместо них можно использовать числа 1991,125; 1991,375; 1991,625; 1991,875; 1992,125;..., которые представляют каждый временной период как срединную точку квартала (добавляя к каждому году 1/8, 3/8, 5/8 и 7/8). Первый квартал 1991 г. представляется своей серединной точкой, 1991,125, которая находится на полпути между началом (1991,000) и концом (1991,250) и может быть вычислена усреднением этих двух крайних точек: (1991,000 + 1991,250)/2 = 1991,125.

Этого можно добиться, воспользовавшись числами 1, 2, 3, ... для представления X непосредственно в виде номера временного периода (квартала или месяца).

В табл. 14.2.6 представлены данные для регрессионного анализа (последние два столбца), цель которого заключается в выявлении долгосрочного тренда в данных Ford Motor Company.

Уравнение регрессии, построенное методом наименьших квадратов, имеет следующий вид:

$$\text{долгосрочный тренд} = 17\,502{,}36 + 535{,}612\ (\text{период времени}).$$

Это выражение указывает на то, что объемы продажи автомобилей Ford Motor Company увеличиваются в среднем на $535,612 (в миллионах) за квартал.

Этот долгосрочный тренд нетрудно прогнозировать, подставляя в уравнение регрессии соответствующий временной период. Например, чтобы найти значение тренда для первого квартала 1998 г., используйте значение $X = 29$, которое будет представлять период времени, следующий за окончанием нашего временного ряда. В этом случае прогноз будет иметь следующий вид:

прогнозируемое значение тренда для первого квартала 1998 г.=
= $17 502,36 + $535,612 (период времени) =
= $17 502,36 + ($535,612 × 29) = $33 035 (в миллионах).

В табл. 14.2.7 представлены прогнозируемые значения (значения долгосрочного тренда и его прогноз на два года вперед по отношению к имеющимся у нас данным). На рис. 14.2.8 показано, как эта линия тренда отражает поведение ряда, учитывающего поправку на сезонные колебания, и продолжается — путем экстраполяции — вправо, определяя прогнозы на будущее с поправкой на сезонные колебания.

Таблица 14.2.6. Объемы продаж автомобилей компании Ford Motor Company с переменными регрессии для вычисления долгосрочного тренда

Год	Квартал	Объем продаж автомобилей, млн дол.	Объем продаж с поправкой на сезонные колебания, Y, млн дол.	Временные периоды, X
1991	1	17 115	16 806	1
1991	2	19 833	18 169	2
1991	3	17 205	19 126	3
1991	4	17 898	18 106	4
1992	1	20 636	20 263	5
1992	2	22 903	20 981	6
1992	3	19 370	21 532	7
1992	4	21 498	21 748	8
1993	1	22 686	22 276	9
1993	2	25 264	23 144	10
1993	3	20 107	22 352	11

Окончание табл. 14.2.6

Год	Квартал	Объем продаж автомобилей, млн дол.	Объем продаж с поправкой на сезонные колебания, *Y*, млн дол.	Временные периоды, *X*
1993	4	23 511	23 785	12
1994	1	26 070	25 599	13
1994	2	28 375	25 994	14
1994	3	24 926	27 708	15
1994	4	27 766	28 089	16
1995	1	28 601	28 084	17
1995	2	29 861	27 355	18
1995	3	24 437	27 165	19
1995	4	27 597	27 918	20
1996	1	28 297	27 786	21
1996	2	31 762	29 097	22
1996	3	26 459	29 413	23
1996	4	31 505	31 872	24
1997	1	30 037	29 494	25
1997	2	32 805	30 052	26
1997	3	28 196	31 343	27
1997	4	31 897	32 268	28



Рис. 14.2.8. Построенная методом наименьших квадратов линия регрессии используется для прогнозирования на основании временного периода ряда объемов продажи с поправкой на сезонные колебания. Эту линию можно продолжить вправо, чтобы получить прогнозы с поправкой на сезонные колебания

Таблица 14.2.7. Объемы продаж автомобилей компании Ford Motor Company и значения долгосрочного тренда

Год	Квартал	Объем продаж автомобилей, млн дол.	Объем продаж с поправкой на сезонные колебания, *Y*, млн дол.	Временные периоды, *X*	Тренд и прогноз с поправкой на сезонные колебания, прогнозируемое значение *Y*, млн дол.
1991	1	17 115	16 806	1	18 038
1991	2	19 833	18 169	2	18 574
1991	3	17 205	19 126	3	19 109
1991	4	17 898	18 106	4	19 645
1992	1	20 636	20 263	5	20 180
1992	2	22 903	20 981	6	20 716
1992	3	19 370	21 532	7	21 252
1992	4	21 498	21 748	8	21 787
1993	1	22 686	22 276	9	22 323
1993	2	25 264	23 144	10	22 858
1993	3	20 107	22 352	11	23 394
1993	4	23 511	23 785	12	23 930
1994	1	26 070	25 599	13	24 465
1994	2	28 375	25 994	14	25 001
1994	3	24 926	27 708	15	25 537
1994	4	27 766	28 089	16	26 072
1995	1	28 601	28 084	17	26 608
1995	2	29 861	27 355	18	27 143
1995	3	24 437	27 165	19	27 679
1995	4	27 597	27 918	20	28 215
1996	1	28 297	27 786	21	28 750
1996	2	31 762	29 097	22	29 286
1996	3	26 459	29 413	23	29 821
1996	4	31 505	31 872	24	30 357
1997	1	30 037	29 494	25	30 893
1997	2	32 805	30 052	26	31 428
1997	3	28 196	31 343	27	31 964
1997	4	31 897	32 268	28	32 500
1998	1			29	33 035
1998	2			30	33 571
1998	3			31	34 106
1998	4			32	34 642

Окончание табл. 14.2.7

Год	Квартал	Объем продаж автомобилей, млн дол.	Объем продаж с поправкой на сезонные колебания, Y, млн дол.	Временные периоды, X	Тренд и прогноз с поправкой на сезонные колебания, прогнозируемое значение Y, млн дол.
1999	1			33	35 178
1999	2			34	35 713
1999	3			35	36 249
1999	4			36	36 784
2000	1			37	37 320
2000	2			38	37 856
2000	3			39	38 391
2000	4			40	38 927

Прогноз: тренд с учетом сезонности

Все, что вам теперь требуется сделать, чтобы прогнозировать будущее, — это учесть сезонность в долгосрочном тренде, вернув ему ожидаемую сезонную вариацию. Для этого достаточно умножить значение тренда на значение сезонного индекса для того периода времени, который вы прогнозируете. Этот процесс является обратным по отношению к внесению поправки на сезонные колебания. Результирующий прогноз включает долгосрочный тренд и сезонную вариацию.

Прогноз = тренд × сезонный индекс

Чтобы предсказать объемы продажи автомобилей компании Ford Motor Company за первый квартал 1998 г., достаточно умножить значение тренда, равное 33 035 (вычисляется с помощью уравнения регрессии для 29-го временного периода), на сезонный индекс для первого квартала, равный 1,0184:

прогноз объема продаж за первый квартал 1998г. =
$33 035 × 1,0184 = $33 643 (в миллионах).

В табл. 14.2.8 представлены прогнозы на три года вперед, по отношению к имеющимся данным. На рис. 14.2.9 показано, как этот учитывающий сезонность тренд отражает рассматриваемый нами ряд и продолжается (путем экстраполяции) вправо, обеспечивая достаточно надежные прогнозы, включающие ожидаемое сезонное поведение объемов продажи.

Стоит ли верить этим прогнозам? Помните, что практически все прогнозы не очень-то достоверны. В конце концов, нерегулярный компонент невозможно предсказать по определению. Кроме того, все эти прогнозы, основанные на тенденциях и сезонных колебаниях, не отражают циклический компонент. Однако их положительная роль заключается хотя бы в том, что они позволяют выявить долгосрочные тенденции нарастания (или убывания), а также повторяющиеся сезонные колебания. Для сравнения можно привести фактические значения объемов продажи за первый и второй кварталы, указанные в отчетах за 1998 год: 29 076 и 31 309.

14.3. Моделирование циклического поведения с помощью ARIMA-процессов Бокса–Дженкинса

Подход Бокса–Дженкинса является одним из лучших методов, позволяющих нам *понять* и *прогнозировать* экономические временные ряды. Результирующие *ARIMA-процессы* представляют собой линейные статистические модели, которые позволяют весьма точно описывать поведение временных рядов самых различных типов, включая даже среднесрочное блуждание так называемого *цикла деловой активности* (или *экономического цикла*). В сравнении с описанным в предыдущем разделе подходом, основанным на трендах и сезонных колебаниях, подход Бокса–Дженкинса отличается более прочным статистическим фундаментом, однако является несколько менее наглядным. После того как вы найдете подходящую модель в рамках семейства ARIMA-процессов Бокса–Дженкинса, в качестве результата можно будет получить вполне приемлемые статистические меры неопределенности (например, стандартную ошибку для прогноза).

Ниже приведен краткий обзор "закулисной" процедуры (т.е. такой, с которой вам не придется иметь дела непосредственно), связанной с использованием методов Бокса–Дженкинса. Этот обзор поможет вам лучше понять, что собой представляют прогнозы и их доверительные интервалы.

1. В семействе ARIMA-процессов Бокса–Дженкинса выбирается достаточно простой процесс, позволяющий получить данные, которые в целом выглядят примерно так же, как ваш ряд (за исключением фактора случайности). Для этого необходимо выбрать конкретный тип модели и оценить требуемые параметры на основе своих данных. Из результирующей модели вы узнаете немало полезного, например, (а) в какой мере каждое наблюдение влияет на будущее и (б) в какой мере каждое наблюдение содержит полезную новую информацию, помогающую прогнозировать будущее.
2. Прогноз на любой момент времени представляет собой ожидаемое (т.е. среднее) будущее значение оцениваемого процесса в этот момент времени.



Рис. 14.2.9. Прогнозирование выполняется путем умножения линии тренда на сезонный индекс. Полученный результат включает тренд и сезонный компонент, но не циклическое и нерегулярное поведение ряда

Представьте себе бесчисленное множество всех допустимых вариантов будущего поведения своего ряда, начиная с ваших исходных данных и экстраполируя эти данные на будущее в соответствии с моделью, выбранной в п. 1. Формула для прогноза позволяет быстро вычислить среднее значение всех этих будущих сценариев.

3. Стандартная ошибка прогноза для любого момента времени представляет собой стандартное отклонение всех возможных (допустимых) будущих значений для этого времени.
4. Границы прогноза простираются выше и ниже прогнозируемого значения так, что (если выбранная вами модель оказалась правильной) с вероятностью, например, 95%, можно утверждать, что будущее значение для любого момента времени уложится в указанные границы прогноза. Эти границы прогноза формируются таким образом, чтобы для каждого будущего периода времени 95%, возможных (и допустимых) вариантов будущего поведения вашего ряда укладывались в эти границы. Сказанное предполагает, что ваш ряд будет и в дальнейшем продолжать вести себя подобно оцениваемому процессу.

Таблица 14.2.8. Объемы и прогнозы продаж автомобилей компании Ford Motor Company

Год	Квартал	Объем продаж автомобилей, млн дол.	Тренд и прогноз с поправкой на сезонные колебания, млн дол.	Сезонный индекс	Тренд с учетом сезонности и прогноз, млн дол.
1991	1	17 115	18 038	1,0184	18 370
1991	2	19 833	18 574	1,0916	20 275
1991	3	17 205	19 109	0,8996	17 190
1991	4	17 898	19 645	0,9885	19 419
1992	1	20 636	20 180	1,0184	20 552
1992	2	22 903	20 716	1,0916	22 614
1992	3	19 370	21 252	0,8996	19 118
1992	4	21 498	21 787	0,9885	21 537
1993	1	22 686	22 323	1,0184	22 734
1993	2	25 264	22 858	1,0916	24 952
1993	3	20 107	23 394	0,8996	21 045
1993	4	23 511	23 930	0,9885	23 654
1994	1	26 070	24 465	1,0184	24 916
1994	2	28 375	25 001	1,0916	27 291
1994	3	24 926	25 537	0,8996	22 972
1994	4	27 766	26 072	0,9885	25 772
1995	1	28 601	26 608	1,0184	27 097
1995	2	29 861	27 143	1,0916	29 630
1995	3	24 437	27 679	0,8996	24 899

Год	Квартал	Объем продаж автомобилей, млн дол.	Тренд и прогноз с поправкой на сезонные колебания, млн дол.	Сезонный индекс	Тренд с учетом сезонности и прогноз, млн дол.
1995	4	27 597	28 215	0,9885	27 890
1996	1	28 297	28 750	1,0184	29 279
1996	2	31 762	29 286	1,0916	31 968
1996	3	26 459	29 821	0,8996	26 827
1996	4	31 505	30 357	0,9885	30 008
1997	1	30 037	30 893	1,0184	31 461
1997	2	32 805	31 428	1,0916	34 307
1997	3	28 196	31 964	0,8996	28 754
1997	4	31 897	32 500	0,9885	32 126
1998	1		33 035	1,0184	33 643
1998	2		33 571	1,0916	36 646
1998	3		34 106	0,8996	30 681
1998	4		34 642	0,9885	34 243
1999	1		35 178	1,0184	35 825
1999	2		35 713	1,0916	38 984
1999	3		36 249	0,8996	32 609
1999	4		36 784	0,9885	36 361
2000	1		37 320	1,0184	38 007
2000	2		37 856	1,0916	41 323
2000	3		38 391	0,8996	34 536
2000	4		38 927	0,9885	38 479

ARIMA-процессы Бокса–Дженкинса представляют собой семейство линейных статистических моделей, основанных на нормальном распределении, которые позволяют имитировать поведение множества различных реальных временных рядов путем комбинирования *процессов авторегрессии*, *процессов интегрирования* и *процессов скользящего среднего* (ARIMA — сокращение от Autoregressive Integrated Moving-Average. — *Прим. ред.*).[10] Результатом является **экономная модель**, т.е. такая, которая использует для описания сложного поведения временного ряда небольшое количество оцениваемых параметров. Несмотря на всю сложность используемой в данном случае теории и формул, сами по себе модели достаточно просты и соответствующие вычисления с помощью компьютера производятся достаточно быстро.

[10] Слово "процесс" относится в этом случае к любой статистической процедуре, которая порождает данные временного ряда.

Сначала мы рассмотрим процесс случайного шума, а затем покажем, как каждый компонент ARIMA-процесса добавляет в модель структуру и гладкость. Мы рассмотрим лишь некоторые основы этих сложных моделей.[11]

Процесс случайного шума не обладает памятью: отправная точка

Процесс случайного шума состоит из случайной выборки (независимых наблюдений) из нормального распределения с постоянным средним и стандартным отклонением. Какие-либо тенденции (тренды) в этом случае отсутствуют, поскольку — по причине независимости — наблюдения не помнят о прошлом поведении ряда.

В соответствии с моделью случайного шума в момент времени t наблюдаемые данные, Y_t, будут состоять из константы, μ (долгосрочное среднее значение процесса), плюс случайный шум, ε_t, с нулевым средним значением.

Процесс случайного шума

Данные = среднее значение + случайный шум.

$$Y_t = \mu + \varepsilon_t .$$

Долгосрочное среднее значение Y равно μ.

Процесс случайного шума в целом носит "плоский" характер — без наклона вверх или вниз. Кроме того, он имеет ярко выраженную тенденцию к нерегулярности и постоянной величине изменчивости, как показано на рис. 14.3.1.

Анализ процесса случайного шума не вызывает особых проблем, поскольку соответствующие данные образуют случайную выборку из нормального распределения — ситуация, уже знакомая вам по материалам глав 9 и 10. Среднее значение является наилучшим прогнозом для любого будущего периода времени, а обычный интервал предсказания для нового наблюдения позволяет получить границы прогноза для любого будущего значения ряда.



Рис. 14.3.1. Процесс случайного шума состоит из независимых наблюдений из нормального распределения. В целом он "плоский" и "дрожащий", с постоянной изменчивостью

[11] Подробнее с этими вопросами можно ознакомиться в книгах Nelson C. R. *Applied Time Series Analysis for Managerial Forecasting* (San Francisco: Holden-Day, 1973); или Box G. E. P. and Jenkins G. M. *Time Series Analysis: Forecasting and Control* (San Francisco: Holden-Day, 1976).

Большинство деловых и экономических данных, имеющих вид временных рядов, помимо компонента случайного шума содержат также определенную структуру. Эту структуру можно представить себе как способ "вспоминания" каждым наблюдением прошлого поведения ряда. Когда такая память сильна, ряд может быть значительно более гладким, чем процесс случайного шума.

Процесс авторегрессии (AR) обладает памятью о своем прошлом

Любое наблюдение **процесса авторегрессии** (часть "AR" названия ARIMA) представляет собой линейную функцию от предыдущего наблюдения плюс случайный шум.[12] Таким образом, процесс авторегрессии помнит о своем предыдущем состоянии и использует эту информацию для определения своего дальнейшего поведения.

В соответствии с моделью процесса авторегрессии в момент времени t значение данных, Y_t, состоит из константы, δ (дельта), плюс коэффициент авторегрессии, φ (фи), умноженный на предшествующее значение данных, Y_{t-1}, плюс случайный шум, ε_t. Обратите внимание, что это модель линейной регрессии, которая прогнозирует текущий уровень ($Y = Y_t$) на основании предшествующего уровня ($X = Y_{t-1}$). В сущности, ряд смещается на величину, пропорциональную $(1 - \varphi)$ своего долгосрочного среднего значения, а затем из этого положения смещается еще на случайное расстояние. Увеличивая φ от 0 к 1, можно придать этому процессу более гладкий вид и сделать его менее похожим на случайный шум.[13] Важно, чтобы коэффициент φ был меньше 1 (по абсолютной величине), что позволит сохранить стабильность процесса.

Процесс авторегрессии

Данные = δ + φ(предыдущее значение) + случайный шум.

$$Y_t = \delta + \varphi Y_{t-1} + \varepsilon_t .$$

Долгосрочное среднее значение Y равно $\delta/(1-\varphi)$.

Поскольку процесс авторегрессии обладает памятью, его значение может какое-то время оставаться высоким, затем какое-то время оставаться низким и т.д. — в результате получается циклическая картина подъемов и снижений по отношению к долгосрочному среднему значению, как показано на рис. 14.3.2. Конкретный процесс, показанный на этом рисунке, характеризуется значением φ, равным 0,8; таким образом, $Y_t = 0{,}8Y_{t-1} + \varepsilon_t$, где ε имеет стандартное отклонение 1 и представляет собой тот же шум, который показан на рис. 14.3.1.

Модели авторегрессии имеют смысл для экономических данных. Они отражают тот факт, что ваши дальнейшие действия в определенной степени зависят

[12] Речь идет о процессе авторегрессии *первого порядка*. Вообще говоря, наблюдение может зависеть от *нескольких* самых последних наблюдений, что определяется множественной регрессией.

[13] Если коэффициент φ меньше нуля, процесс авторегрессии может фактически оказаться даже *более* неровным, чем случайный шум, поскольку он стремится к тому, чтобы попеременно принимать высокие и низкие значения. Будем предполагать, что коэффициент φ положителен; в этом случае процесс авторегрессии оказывается более гладким, чем случайный шум.

от вашего текущего состояния (что выражается коэффициентом авторегрессии, φ) и в определенной степени — от того, что происходит с вами по ходу дела (что выражается компонентом случайного шума).

Прогнозирование с помощью процесса авторегрессии выполняется на основе прогнозируемых значений из оцениваемого уравнения регрессии при продвижении вперед на единицу времени; таким образом, прогнозируемое значение Y_{t+1} равно $\delta + \varphi Y_t$. ("Шляпки" над коэффициентами указывают, что эти коэффициенты являются оценками, полученными на основе данных, а не значениями генеральной совокупности.) Полученный прогноз является неким компромиссом между самым последним значением данных и долгосрочным средним значением ряда. Чем дальше в будущее вы пытаетесь заглянуть, тем ближе окажется ваш прогноз к оценке значения долгосрочного среднего, поскольку процесс постепенно "забывает" свое отдаленное прошлое.

Процесс скользящего среднего (MA) имеет ограниченную память

Любое наблюдение **процесса скользящего среднего** (moving-average process — "MA" в ARIMA) состоит из константы, μ (долгосрочное среднее значение процесса), плюс независимый случайный шум минус часть предыдущего случайного шума.[14] Процесс скользящего среднего не помнит в точности своего прошлого, но помнит компонент случайного шума того состояния, в котором он (процесс) находился. Таким образом, его память ограничена одним шагом в будущее; за пределами этого шага для процесса все начинается заново.

В соответствии с моделью процесса скользящего среднего в момент времени t значение данных, Y_t, состоит из константы, μ, плюс случайный шум, ε_t, минус некоторая доля, θ (тета, коэффициент скользящего среднего), предыдущего случайного шума. Уменьшая этот коэффициент θ от 0 до −1, этот процесс можно сделать несколько менее похожим на случайный шум, однако он становится лишь чуть более гладким.[15]

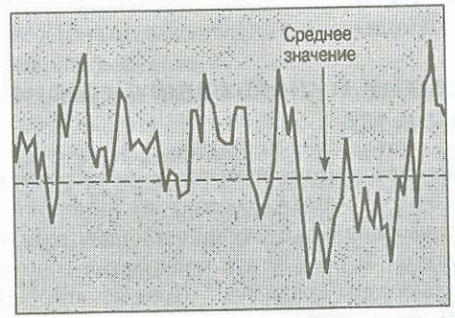
Процесс случайного среднего

Данные = μ + случайный шум − θ(предыдущий случайный шум).

$$Y_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1}.$$

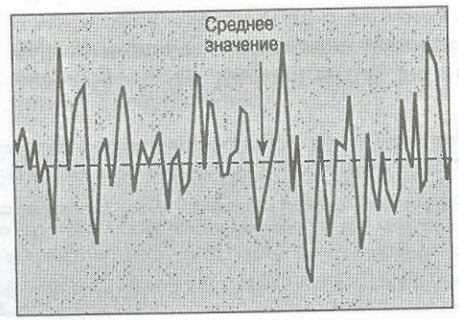
Долгосрочное среднее значение Y равно μ.

Поскольку процесс случайного среднего обладает памятью, он позволяет получать смежные пары наблюдений, которые *оба* весьма вероятно оказываются либо высокими, либо низкими. Однако поскольку память этого процесса носит ограниченный характер, соответствующий ряд снова оказывается случайным уже после двух шагов. В результате получается ряд, который не является в такой же степени случайным, как ряд чистого случайного шума. Сравнивая процесс скользящего среднего, показанный на рис. 14.3.3, с рядом чистого случайного шума, показанным на рис. 14.3.1, можно заметить снижение случайности. Конкретный процесс, показанный на рис. 14.3.3, характеризуется значением θ, равным −0,8; таким образом, $Y_t = \varepsilon_t + 0{,}8\varepsilon_{t-1}$, где ε имеет стандартное отклонение 1 и представляет собой такой же шум, как показанный на рис. 14.3.1. Фактически этот ряд представляет собой скользящее среднее случайного шума.



Процесс авторегрессии (AR)

Рис. 14.3.2. Процесс авторегрессии проявляется как уравнение линейной регрессии, в котором текущее значение помогает прогнозировать следующее значение. Обратите внимание, что этот ряд оказывается менее неровным, чем чистый шум (сравните с рис. 14.3.1), и что он может отклоняться от своего долгосрочного среднего значения в течение продолжительных периодов времени



Процесс скользящего среднего (MA)

Рис. 14.3.3. Процесс скользящего среднего запоминает лишь часть предыдущего шума. Результат оказывается несколько менее нерегулярным, чем чисто случайный шум (сравните с рис. 14.3.1). Спустя два периода снова становится случайным, поскольку он не помнит, где он был

[14] Речь идет о процессе скользящего среднего *первого порядка*. Вообще говоря, наблюдение может зависеть от нескольких самых последних компонентов случайного шума, и ограниченная память может охватывать несколько шагов.

[15] Если коэффициент θ положителен, процесс скользящего среднего может фактически оказаться даже несколько *более* неровным, чем случайный шум, поскольку он стремится к тому, чтобы попеременно принимать высокие и низкие значения. Будем предполагать, что коэффициент θ отрицателен; в этом случае процесс скользящего среднего оказывается несколько более гладким, чем случайный шум.

Чистые модели скользящего среднего имеют лишь ограниченное применение для экономических данных вследствие своей ограниченной памяти (на что указывает коэффициент скользящего среднего, θ). Их лучше всего использовать в сочетании с авторегрессионными процессами, что дает возможность сосредоточить внимание на самых последних событиях (в отличие от того, что позволяют процессы чистой авторегрессии).

Прогнозирование следующего наблюдения с помощью процесса скользящего среднего основывается на оценке текущего случайного шума, ε_t. За пределами следующего наблюдения наилучшим прогнозом является оценка долгосрочного среднего значения, μ, поскольку этот процесс забывает все, за исключением самого последнего, свое прошлое.

Процесс авторегрессии и скользящего среднего (ARMA) сочетает в себе AR и MA

Любое наблюдение **процесса авторегрессии и скользящего среднего** (autoregressive moving-average process — ARMA) состоит из линейной функции от предыдущего наблюдения плюс независимый случайный шум минус некоторая доля предыдущего случайного шума. В результате получается сочетание процесса авторегрессии с процессом скользящего среднего.[16] Процесс авторегрессии и скользящего среднего запоминает как свое предыдущее состояние, так и компонент случайного шума предыдущего состояния. Таким образом, его память сочетает в себе память процесса авторегрессии с памятью процесса скользящего среднего. В результате получается процесс авторегрессии с улучшенной краткосрочной памятью.

В соответствии с моделью процесса авторегрессии и скользящего среднего (ARMA) в момент времени t значение данных, Y_t, состоит из константы, δ, плюс коэффициент авторегрессии, φ, умноженный на предшествующее значение данных, Y_{t-1}, плюс случайный шум, ε_t, минус некоторая доля, θ, предыдущего случайного шума. Это напоминает модель линейной регрессии за исключением того, что соответствующие ошибки не являются независимыми. По мере изменения φ от 0 к 1 и по мере изменения θ от 0 к −1 результирующий процесс принимает более гладкий вид и меньше напоминает случайный шум. С точки зрения стабильности процесса важно, чтобы коэффициент φ был меньше 1 (по своему абсолютному значению).

Процесс авторегрессии и скользящего среднего (ARMA)

Данные = δ + φ(предыдущее значение) + случайный шум − θ(предыдущий случайный шум).

$$Y_t = \delta + \varphi Y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}.$$

Долгосрочное среднее значение Y равно $\delta/(1-\varphi)$.

Вследствие наличия у ARMA-процесса памяти его значение может какое-то время оставаться высоким, затем какое-то время оставаться низким и т.д. — в результате получается циклическая картина подъемов и снижений по отношению к

[16] Речь идет о процессе авторегрессии и скользящего среднего *первого порядка*. Вообще говоря, наблюдение может зависеть от нескольких самых последних наблюдений и компонентов случайного шума.

долгосрочному среднему значению, как показано на рис. 14.3.4. Конкретный процесс, показанный на этом рисунке, характеризуется значением φ, равным 0,8, и значением θ, равным −0,8; таким образом, $Y_t = 0{,}8Y_{t-1} + \varepsilon_t + 0{,}8\varepsilon_{t-1}$, где ε имеет стандартное отклонение 1 и представляет собой такой же шум, как и показанный на рис. 14.3.1. Поскольку процесс чистой авторегрессии (рис. 14.3.2) содержит точно такой же случайный шум, сравнение рис. 14.3.2 и 14.3.4 демонстрирует вклад составляющей скользящего среднего в гладкость этого ARMA-процесса.

Сочетание процесса авторегрессии с процессом скользящего среднего доказало свою высокую эффективность в применении к экономическим данным. Подбирая значения коэффициентов φ и θ, можно выбрать такую модель, которая подойдет для любой совокупности данных циклических и нерегулярных временных рядов, несмотря на огромное их разнообразие.

Прогнозирование следующего наблюдения с помощью ARMA-процесса выполняется путем комбинирования прогнозируемого значения из оцениваемого уравнения авторегрессии (прогнозируемое Y_{t+1} равно $\delta + \varphi Y_t$, причем "шляпки" над коэффициентами в этом случае также указывают оценки) с оценкой текущего случайного шума, ε_t. За пределами следующего наблюдения наилучший прогноз основывается только на предыдущем прогнозируемом значении. Чем дальше в будущее вы пытаетесь заглянуть, тем ближе оказывается ваш прогноз к оцениваемому долгосрочному среднему значению, поскольку процесс постепенно "забывает" свое отдаленное прошлое.

Пример. *Прогнозирование уровня безработицы с помощью ARMA-процесса*

В табл. 14.3.1 представлен уровень безработицы среди гражданских работников в США. Процент безработных фиксировался ежегодно с 1960 по 1998 г. включительно. Соответствующий график представлен на рис. 14.3.5.

Оценка ARMA-модели была выполнена для этой совокупности данных методом наименьших квадратов, и соответствующие результаты представлены в табл. 14.3.2.[17] Обратите внимание, что и коэффициент авторегрессии, и коэффициент скользящего среднего являются статистически значимыми, о чем свидетельствуют соответствующие *p*-значения *t*-отношения.

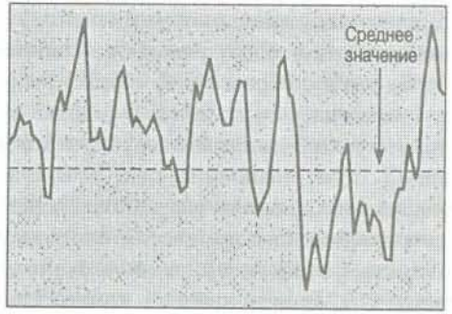
Эти результаты дают нам ARMA-модель, которая позволяет получить временной ряд данных, похожий на наши данные об уровне безработицы — с тем же типом нерегулярности, гладкости и циклического поведения. Эта оценка ARMA-модели имеет следующий вид:

данные = 2,070 + 0,649(предыдущее значение) + случайный шум + 0,402(предыдущий случайный шум);

$$Y_t = 2{,}070 + 0{,}649Y_{t-1} + \varepsilon_t + 0{,}402\varepsilon_{t-1}.$$

В соответствии с этой оценкой модели за год процент безработных изменяется на небольшую величину, поскольку значение данных за каждый год определяется с учетом уровня предыдущего года. Точнее говоря, чтобы найти значение данных за каждый год, необходимо передвинуть текущий уровень безработицы на (1 – 0,649) = 35,1% по направлению к долгосрочному среднему значению 5,900, затем добавить новый случайный шум и, наконец, добавить 40,2% предыдущего случайного шума.

[17] Цель получения оценки методом наименьших квадратов заключается в минимизации компонента шума, чтобы компонент авторегрессии и компонент скользящего среднего используемой модели как можно в большей степени выявили структуру ряда. Когда шум представляет собой случайную выборку из нормального распределения, мощный общий метод максимального подобия дает те же оценки, что и метод наименьших квадратов (причиной этого является наличие экспоненциального квадратного члена в функции нормальной плотности).



Процесс авторегрессии и скользящего среднего (ARMA)

Рис. 14.3.4. В процессе авторегрессии и скользящего среднего (ARMA) и текущее значение, и текущий шум помогают определить следующее значение. Полученный результат оказывается более гладким благодаря памяти процесса авторегрессии в сочетании с дополнительной краткосрочной (на один шаг вперед) памятью процесса скользящего среднего

Насколько данные, полученные из оценки ARMA-процесса, соответствуют реальному уровню безработицы? На рис. 14.3.6 показаны фактический процент безработных и два варианта имитационного моделирования, полученных с помощью оценки ARMA-процесса. В этих двух случаях использовался один и тот же начальный уровень безработицы (5,4% за 1960 г.), но разный случайный шум. Эти два варианта имитационного моделирования можно рассматривать как альтернативные сценарии того, что могло бы произойти.

Главная цель анализа временных рядов в сфере бизнеса — прогнозирование. В табл. 14.3.3 представлены прогнозы процента безработных (вместе с границами прогноза) на последующие десять лет. Эти прогнозы вычислены на основе оценки ARMA-модели. Рис. 14.3.7 демонстрирует, что прогноз постепенно продвигается от последнего значения ряда (4,5% за 1998 г.) в направлении долгосрочного среднего значения (5,9%). Этот прогноз, лучший из тех, который можно было сделать, основываясь лишь на данных из табл. 14.3.1 и использованной нами ARMA-модели, говорит о том, что в среднем мы ожидаем, что ряд постепенно "забудет" о своем пребывании выше линии долгосрочного среднего значения и в конце концов вернется к этому значению. Разумеется, мы также ожидаем, что он продолжит свое циклическое и нерегулярное поведение (именно это является причиной того, что наши 95% границы прогноза столь широки).

На рис. 14.3.8 показаны три варианта моделирования будущего, созданные на основе оценки ARMA-модели с использованием нового, независимого шума. Полученный прогноз представляет среднее всех таких результатов моделирования будущего. Границы прогноза включают средние 95% всех таких результатов моделирования в каждый период времени будущего.

Таблица 14.3.1. Процент безработных среди гражданских лиц

Год	Процент безработных	Год	Процент безработных	Год	Процент безработных
1960	5,4	1973	4,9	1986	7,0
1961	6,8	1974	5,5	1987	6,2
1962	5,6	1975	8,5	1988	5,5
1963	5,7	1976	7,7	1989	5,2
1964	5,2	1977	7,1	1990	5,6

Окончание табл. 14.3.1

Год	Процент безработных	Год	Процент безработных	Год	Процент безработных
1965	4,6	1978	6,1	1991	6,8
1966	3,8	1979	5,8	1992	7,5
1967	3,8	1980	7,2	1993	6,9
1968	3,6	1981	7,5	1994	6,2
1969	3,5	1982	9,6	1995	5,6
1970	4,9	1983	9,7	1996	5,4
1971	5,9	1984	7,5	1997	5,0
1972	5,6	1985	7,2	1998*	4,5

*Показатель за 1998 г. соответствует первым девяти месяцам.

Приведенные данные представляют собой среднегодовые значения помесячных показателей, источником которых является Бюро трудовой статистики США, `http://stats.bls.gov/webapps/legacy/cpsatab5.htm`, осень 1998 г.



Таблица 14.3.2. Оценки ARMA-модели, полученные на основе данных об уровне безработицы

Коэффициент	Оценка	Стандартная ошибка	t-отношение	p
Авторегрессия ($\hat{\varphi}$)	0,649	0,156	4,15	0,000
Скользящее среднее ($\hat{\theta}$)	–0,402	0,183	–2,20	0,035
Константа ($\hat{\delta}$)	2,070	0,205	10,11	0,000
Среднее ($\hat{\delta}/(1-\hat{\varphi})$)	5,900	0,584	10,11	0,000
Стандартное отклонение случайного шума	0,907			

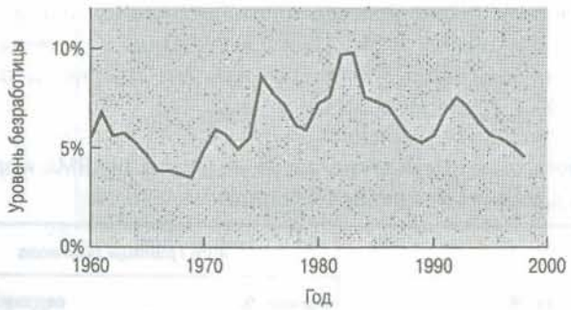


Рис. 14.3.5. Уровень безработицы в США с 1960 по 1998 гг. Обратите внимание на степень гладкости (совершенно очевидно, что это не просто случайный шум) и тенденцию к цикличности поведения



Рис. 14.3.6. Два варианта моделирования на основании оценки ARMA-процесса показаны вместе с фактическими уровнями безработицы. Обратите внимание, как это искусственное моделирование в целом довольно точно отражает характер реальных данных — с точки зрения гладкости, нерегулярности и наличия циклов. Эта способность моделей вести себя подобно реальным рядам является важной особенностью анализа Бокса–Дженкинса.

Чистый интегрированный (I) процесс помнит, где он находился, и затем движется случайно

Каждое наблюдение **чистого интегрированного (I) процесса** (pure integrated (I) process), называемого также **случайным блужданием**, заключается в случайном шаге в сторону от текущего наблюдения. Этот процесс знает, где он находится, но забыл, как он попал туда. Случайное блуждание иногда называют **нестационарным процессом**, поскольку с течением времени он имеет тенденцию уходить все дальше от той точки, в которой он находился. В отличие от нестационарного процесса, модель авторегрессии, модель скользящего среднего и ARMA-модель представляют собой **стационарные процессы**, поскольку на протяжении длительных периодов времени они, как правило, ведут себя сходным образом, оставаясь в относительной близости от своего долгосрочного среднего значения.

В соответствии с моделью случайного блуждания в момент t значение данных, Y_t, состоит из константы, δ (составляющая "дрейфа"), плюс предыдущее значение данных, Y_{t-1}, плюс случайный шум, ε_t.

Таблица 14.3.3. Прогнозы и границы прогнозов, задаваемые ARMA-моделью, оцененной на данных об уровне безработицы

Год	Прогноз, %	95% границы прогноза	
		нижняя, %	верхняя, %
1999	4,78	2,71	6,84
2000	5,06	2,37	7,76
2001	5,26	2,32	8,20

Год	Прогноз, %	95% границы прогноза	
		нижняя, %	верхняя, %
2002	5,40	2,35	8,44
2003	5,49	2,39	8,58
2004	5,55	2,43	8,67
2005	5,59	2,46	8,72
2006	5,62	2,49	8,75
2007	5,64	2,50	8,78
2008	5,65	2,51	8,79



Рис. 14.3.7. Процент безработицы, его прогноз на десять лет вперед и 95% границы прогноза, вычисленные на основании оценки ARMA-модели. Этот прогноз свидетельствует о том, что рассматриваемый ряд в среднем будет постепенно забывать о том, что он находится ниже линии долгосрочного среднего значения. Границы прогноза достаточно широки, чтобы предвидеть будущее циклическое и нерегулярное поведение

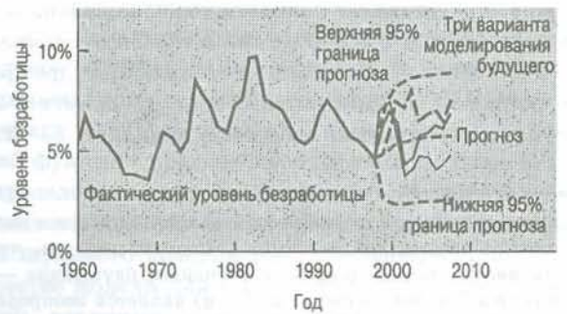


Рис. 14.3.8. Процент безработицы, его прогноз на десять лет вперед, 95% границы прогноза и три варианта моделирования будущего. Прогноз дает среднее значение всех таких вариантов моделирования для каждого из будущих периодов времени. Границы прогноза включают 95% всех таких результатов моделирования для каждого периода времени в будущем

Несмотря на то что эта модель очень похожа на модель авторегрессии при $\varphi = 1$, поведение этих двух моделей существенно разное.[18] Составляющая дрейфа, δ, позволяет нам заставить процесс блуждать случайным образом, смещаясь со временем в среднем несколько вверх (если $\delta > 0$) или вниз (если $\delta < 0$). Однако, даже если $\delta = 0$, у исследователя может *создаться впечатление*, что ряд с течением времени проявляет тенденции в сторону возрастания и в сторону убывания.

Чистый интегрированный процесс (случайное блуждание)

Данные = δ + предыдущее значение + случайный шум.

$$Y_t = \delta + Y_{t-1} + \varepsilon_t .$$

Нельзя рассчитывать на то, что с течением времени Y будет оставаться достаточно близким к какому-либо долгосрочному среднему значению.

Самый простой способ проанализировать чистый интегрированный процесс — воспользоваться рядом *разностей*, $Y_t - Y_{t-1}$, которые соответствуют процессу случайного шума.[19]

Чистый интегрированный процесс (случайное блуждание) в форме разностей

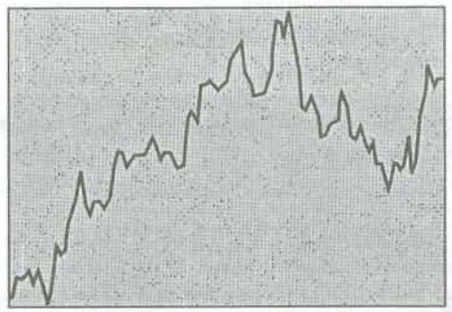
Данные – предыдущее значение = δ + случайный шум.

$$Y_t - Y_{t-1} = \delta + \varepsilon_t .$$

Поскольку тенденция возврата к долгосрочному среднему значению в данном случае не наблюдается, случайные блуждания могут вводить в заблуждение, создавая впечатление наличия тенденций там, где на самом деле их нет. Случайное блуждание, показанное на рис. 14.3.9, было получено при $\delta = 0$, поэтому *никаких реальных тенденций здесь нет* — лишь случайные изменения. Этому ряду ничего "неизвестно" о том, когда он достиг своей наивысшей точки, — где бы это ни произошло, он просто продолжает свои случайные блуждания, как делал и до того. Был использован такой же случайный шум, как и на рис. 14.3.1, который, в свою очередь, представляет разности ряда, показанного на рис. 14.3.9.

Прогнозирование следующего наблюдения с помощью случайного блуждания осуществляется путем добавления оценки составляющей дрейфа, δ, к текущему наблюдению. Для каждого дополнительного периода времени, который вы экстраполируете в будущее, добавляется очередное значение δ. Если составляющая дрейфа отсутствует (т.е. если вы полагаете, что $\delta = 0$), тогда прогнозом всех будущих значений является текущее значение. Границы прогноза в любом из этих случаев с течением времени будут продолжать расширяться (больше, чем в случае ARMA-процессов) вследствие нестационарности.

Модель случайных блужданий очень важна сама по себе (например, в качестве модели фондовой биржи). Кроме того, она является важным строительным блоком, когда используется в сочетании с ARMA-моделями для создания ARIMA-моделей. Это добавляет гибкости и позволяет анализировать более сложные временные ряды.



Интегрированный (I) процесс,
или случайное блуждание без дрейфа

Рис. 14.3.9. Чистый интегрированный (I) процесс, или случайное блуждание, с нулевой составляющей дрейфа может создать впечатление наличия тенденций, в то время как на самом деле никаких тенденций нет. Ряд помнит лишь то, где он находится, а его последующие шаги носят исключительно случайный характер

Процесс авторегрессионного интегрированного скользящего среднего (ARIMA) помнит свои изменения

Если изменения или разности в ряде вырабатываются процессом авторегрессии и скользящего среднего (ARMA), то сам этот ряд соответствует **процессу авторегрессионного интегрированного скользящего среднего (ARIMA)** (autoregressive integrated moving-average (ARIMA) process). Таким образом, *изменение* в процессе состоит из линейной функции предыдущего изменения плюс независимый случайный шум минус определенная доля предыдущего случайного шума.[20] Этот процесс знает, где он находится, помнит, как он попал в это состояние, и помнит даже часть предыдущего шумового компонента. Следовательно, ARIMA-процесс можно использовать в качестве модели для совокупностей данных временного ряда, которые являются очень гладкими, с медленными изменениями направления. Эти ARIMA-процессы являются нестационарными из-за включения в них интегриро-

[18] Вот почему в определении процесса авторегрессии и ARMA-процесса мы решили ограничить φ значениями, меньшими по абсолютной величине 1. Следует помнить, что модель авторегрессии и ARMA-модель являются стационарными, а случайное блуждание — нет. При $\varphi = 1$ для ARMA-процесса долгосрочное среднее значение $\delta/(1 - \varphi)$ является неопределенным (по причине деления на нуль).

[19] В случае фондовой биржи и некоторых других совокупностей данных экономического характера более целесообразным может оказаться использование *процентных изменений*, $(Y_t - Y_{t-1})/Y_{t-1}$. Это одна из разновидностей идеи, связанной с использованием дифференциалов. Если говорить более определенно, то пользоваться процентными изменениями удобно в тех случаях, когда закону случайных блужданий (с относительно небольшими шагами) подчиняются *логарифмы* данных.

[20] Речь идет о процессе авторегрессионного интегрированного скользящего среднего *первого порядка*. Вообще говоря, изменение может зависеть от нескольких последних изменений и компонентов случайного шума.

ванного компонента. Таким образом, с течением времени подобный ряд, как правило, уходит все дальше и дальше от своего исходного состояния.

В соответствии с моделью процесса авторегрессионного интегрированного скользящего среднего в момент t изменение значения данных, $Y_t - Y_{t-1}$, состоит из константы, δ, плюс коэффициент авторегрессии, φ, умноженный на предыдущее изменение, $Y_{t-1} - Y_{t-2}$, плюс случайный шум, ε_t, минус определенная доля, θ, предыдущего случайного шума. Это напоминает модель линейной регрессии, записанную в *разностях*, за исключением того, что ошибки не являются независимыми. По мере того как φ изменяется от 0 до 1, а θ — от 0 до − 1, результирующий процесс становится все более гладким. Важно, чтобы коэффициент φ был меньше 1 (по абсолютному значению) — в этом случае обеспечивается стабильность процесса (разностного).

Процесс авторегрессионного интегрированного скользящего среднего (ARIMA) в разностной форме

Изменение данных = δ + φ(предыдущее значение) + случайный шум − θ(предыдущий случайный шум).

$$Y_t - Y_{t-1} = \delta + \varphi(Y_{t-1} - Y_{t-2}) + \varepsilon_t - \theta\varepsilon_{t-1}.$$

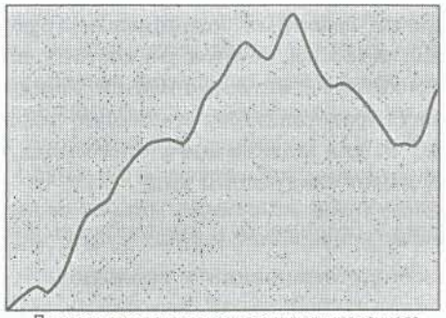
Долгосрочное среднее значение изменения в Y равно $\delta/(1-\varphi)$. Нельзя рассчитывать, что с течением времени Y останется достаточно близким к какому-либо долгосрочному среднему значению.

На рис. 14.3.10 представлен ARIMA-процесс, полученный путем суммирования (иногда говорят *интегрирования*) ARMA-процесса, показанного на рис. 14.3.4. Поскольку он включает тот же случайный шум, что и в случайном блуждании, представленном на рис. 14.3.9, можно заметить, как компонент авторегрессии и компонент скользящего среднего сглаживают изменения, в то же время сохраняя — в целом — поведение соответствующего ряда.

Прогнозирование в случае ARIMA-модели осуществляется путем прогнозирования изменений ARMA-модели для разностей. Вследствие нестационарности такие прогнозы могут проявлять тенденцию к бесконечному нарастанию (или снижению), а границы прогнозов будут расширяться по мере распространения этих прогнозов на все более отдаленное будущее.

В каких случаях полезен такой переход к разностям? ARIMA-модель (в разностях) будет полезна в тех ситуациях, в которых нет тенденции возврата к долгосрочному среднему значению (например, биржевая цена, валовой национальный продукт США, индекс потребительских цен или объем продаж в вашей фирме). ARMA-модель (в которой не используются разности) будет полезна в ситуациях, в которых ряд стремится оставаться вблизи долгосрочного среднего значения (примерами подобных рядов является уровень безработицы, процентные ставки, изменения индекса цен, а также отношение суммарных обязательств или долгосрочных заимствований к акционерному капиталу вашей фирмы).

Возможно создание и более совершенных ARIMA-моделей, которые включали бы сезонное поведение поквартальных и помесячных рядов. Основная идея состоит в том, чтобы включить в уравнения модели, помимо значения за последний месяц, еще и значение за последний год.



Процесс авторегрессионного интегрированного скользящего среднего (ARIMA)

Рис. 14.3.10. Процесс авторегрессионного интегрированного скользящего среднего (ARIMA) помнит, где он находится, как он попал в это состояние, а также определенную часть предыдущего шума. В результате получается модель очень гладкого временного ряда. Сравните его со случайным блужданием (с тем же шумом), показанным на рис. 14.3.9

14.4. Дополнительный материал

Резюме

Временной ряд отличается от данных об одном временном срезе в том отношении, что в случае временных рядов сама последовательность наблюдений несет в себе важную информацию. Чтобы методами, описанными в предыдущих главах (например, доверительные интервалы и проверки гипотез), можно было пользоваться и в случае временных рядов, их необходимо определенным образом модифицировать, поскольку временной ряд, как правило, не является случайной выборкой из некоторой генеральной совокупности.

Главная цель анализа временных рядов заключается в создании прогнозов, т.е. *прогнозировании будущего*. Эти прогнозы основываются на той или иной **модели** (которую также называют **математической моделью**, или **процессом**). Модель представляет собой систему уравнений, которая позволяет получить некий набор искусственных совокупностей данных, относящихся к категории временных рядов. **Прогноз** представляет собой ожидаемое (т.е. среднее) значение будущего поведения оцениваемой модели. Подобно всем оценкам, прогнозы обычно не соответствуют действительности. **Границами прогноза** являются доверительные границы прогноза (если используемая модель позволяет определять их); если используемая модель корректна по отношению к исследуемым данным, то будущее наблюдение с вероятностью, например, 95% попадет в эти границы.

Анализ трендов и сезонных колебаний представляет собой непосредственный, интуитивный подход к оцениванию четырех базовых компонентов помесячных или поквартальных временных рядов: долгосрочного тренда (тенденции), сезонных колебаний (сезонности), циклической вариации и нерегулярного компонента. Долгосрочный **тренд** указывает *действительно* долгосрочное поведение временного ряда — как правило, в виде прямой линии или экспоненциальной кри-

вой. Точно повторяющийся **сезонный компонент** определяет влияние времени года. Среднесрочный **циклический компонент** состоит из последовательных повышений и понижений, которые *не* повторяются каждый год. Краткосрочный **нерегулярный компонент** представляет остаточную вариацию, которую невозможно объяснить. Формула для модели временного ряда, основанная на трендах и сезонных колебаниях, имеет следующий вид:

$$\text{данные} = \text{тренд} \times \text{сезонность} \times \text{цикличность} \times \text{нерегулярность}.$$

При использовании метода **отношения к скользящему среднему** значения ряда делят на значение гладко скользящего среднего следующим образом:

1. **Скользящее среднее** представляет собой новый ряд, созданный путем усреднения соседних наблюдений. Для каждого усреднения мы используем данные за целый год, что дает возможность избавиться от влияния сезонного компонента.
2. Скользящее среднее = тренд × цикличность.
3. Деление исходного ряда на сглаженный ряд скользящего среднего дает нам отношение к скользящему среднему, которое включает как сезонные, так и нерегулярные значения. Проводя группирование по времени года, а затем выполняя усреднение в полученных группах, получаем **сезонный индекс** для каждого времени года. Сезонный индекс указывает, насколько больше (или меньше) бывает анализируемый показатель в соответствующий период времени по сравнению с типичным периодом, характерным для года в целом. **Поправка на сезон** позволяет избавиться от ожидаемого сезонного компонента в наблюдении (путем деления ряда на сезонный индекс для соответствующего периода времени) и дает возможность выявлять скрытые тенденции, непосредственно сравнивая один квартал или месяц с другим (после внесения поправки на сезон).

$$\text{Сезонность} \times \text{Нерегулярность} = \frac{\text{Данные}}{\text{Скользящее среднее}}$$

$$\text{Сезонный индекс} =$$

$$= \text{Среднее значение} \left(\frac{\text{Данные}}{\text{Скользящее среднее}} \right) \text{ за соответствующий сезон}$$

$$\text{Значение с поправкой на сезон} = \left(\frac{\text{Данные}}{\text{Сезонный индекс}} \right) =$$

$$= \text{Тренд} \times \text{Цикличность} \times \text{Нерегулярность}.$$

4. Регрессия ряда с поправкой на сезон (Y) по времени (X) используется для оценивания долгосрочного тренда в виде прямой линии как функции от времени и дает возможность получить прогноз с поправкой на сезон. Этой возможностью удобно пользоваться лишь в случае, если долгосрочный тренд в вашем ряде является линейным.
5. Прогнозирование можно выполнять с помощью *внесения сезонности в тренд*, т.е. путем умножения тренда на соответствующий сезонный индекс.

Процессы Бокса–Дженкинса образуют семейство линейных статистических моделей, основанных на нормальном распределении. Гибкость этого семейства моделей позволяет имитировать поведение множества различных реальных временных рядов путем комбинирования *процессов авторегрессии* (autoregressive — AR), *интегрированных процессов* (integrated — I) и *процессов скользящего среднего* (moving-average — MA). Результатом является **экономная модель**, которая использует для описания сложного поведения временного ряда небольшое количество оцениваемых параметров. Ниже приведен краткий обзор используемых для этого шагов.

1. В семействе ARIMA-процессов Бокса–Дженкинса выбирается достаточно простой процесс, позволяющий получить данные, которые в целом выглядят примерно так же, как ваш ряд (за исключением фактора случайности).
2. Прогноз на любой момент времени представляет собой ожидаемое (т.е. среднее) будущее значение оцениваемого процесса в этот момент времени.
3. Стандартная ошибка прогноза для любого периода времени представляет собой стандартное отклонение будущего значения оцениваемого процесса в этот период времени.
4. Границы прогноза простираются выше и ниже прогнозируемого значения; таким образом, с вероятностью, например, 95% можно утверждать, что будущее значение для любого периода времени будет находиться в указанных границах прогноза. Сказанное предполагает, что ваш ряд будет и в дальнейшем вести себя подобно оцениваемому процессу.

Процесс случайного шума состоит из случайной выборки (независимых наблюдений) из нормального распределения с постоянным средним и стандартным отклонением. Это среднее значение является наилучшим прогнозом для любого будущего периода времени, а обычный интервал прогнозирования для нового наблюдения позволяет получить границы прогноза для любого будущего значения ряда. Формула для процесса случайного шума имеет следующий вид:

$$\text{данные} = \text{среднее значение} + \text{случайный шум};$$
$$Y_t = \mu + \varepsilon_t;$$

долгосрочное среднее значение Y равно μ.

Любое наблюдение **процесса авторегрессии** состоит из линейной функции от предыдущего наблюдения плюс случайный шум. Прогнозирование с помощью процесса авторегрессии выполняется на основе прогнозируемых значений из оценки уравнения регрессии $\delta + \varphi Y_t$. Полученный прогноз является неким компромиссом между самым последним значением данных и долгосрочным средним значением ряда. Чем дальше в будущее вы пытаетесь заглянуть, тем ближе окажется ваш прогноз к долгосрочному среднему значению. Соответствующая формула имеет следующий вид:

$$\text{данные} = \delta + \varphi(\text{предыдущее значение}) + \text{случайный шум};$$
$$Y_t = \delta + \varphi Y_{t-1} + \varepsilon_t;$$

долгосрочное среднее значение Y равно $\delta/(1-\varphi)$.

Любое наблюдение **процесса скользящего среднего** состоит из константы, μ (долгосрочное среднее значение процесса) плюс независимый случайный шум минус некоторая часть предыдущего случайного шума:

$$\text{данные} = \mu + \text{случайный шум} - \theta(\text{предыдущий случайный шум});$$
$$Y_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1}.$$

Долгосрочное среднее значение Y равно μ. В результате получается скользящее среднее двух наблюдений в некоторый момент времени из процесса случайного шума. Прогнозирование следующего наблюдения основывается на оценке текущего случайного шума, ε_t; за этими границами наилучшим прогнозом является оцениваемое долгосрочное среднее значение.

Любое наблюдение **процесса авторегрессии и скользящего среднего** (autoregressive moving-average process — ARMA) состоит из линейной функции от предыдущего наблюдения плюс независимый случайный шум минус некоторая доля предыдущего случайного шума. В результате получается сочетание процесса авторегрессии с процессом скользящего среднего:

$$\text{данные} = \delta + \varphi(\text{предыдущее значение}) + \text{случайный шум} -$$
$$- \theta(\text{предыдущий случайный шум});$$
$$Y_t = \delta + \varphi Y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}.$$

Долгосрочное среднее значение Y равняется $\delta/(1-\varphi)$. Прогнозирование следующего наблюдения выполняется путем комбинирования прогнозируемого значения из оценки уравнения авторегрессии $\delta + \varphi Y_t$ с оценкой текущего случайного шума, ε_t. За пределами следующего наблюдения наилучший прогноз основывается только на предыдущем прогнозируемом значении. Чем дальше в будущее вы пытаетесь заглянуть, тем ближе оказывается ваш прогноз к оцениваемому долгосрочному среднему значению.

Каждое наблюдение **чистого интегрированного (I) процесса**, называемого также **случайным блужданием**, представляет собой случайный шаг в сторону от текущего наблюдения. Случайное блуждание иногда называют **нестационарным процессом**, поскольку такой процесс имеет тенденцию с течением времени уходить все дальше от той точки, в которой он находился. В отличие от этого модель авторегрессии, модель скользящего среднего и ARMA-модель представляют собой **стационарные процессы**, поскольку на протяжении длительных периодов времени они, как правило, ведут себя сходным образом, оставаясь в относительной близости от своего долгосрочного среднего значения. Прогнозирование следующего наблюдения с помощью случайного блуждания осуществляется путем добавления к текущему наблюдению оценки составляющей дрейфа, δ, для каждого дополнительного периода в будущем. Чистый интегрированный процесс (случайное блуждание) можно представить в виде следующего выражения:

$$\text{данные} = \delta + \text{предыдущее значение} + \text{случайный шум};$$
$$Y_t = \delta + Y_{t-1} + \varepsilon_t.$$

Нельзя рассчитывать на то, что с течением времени Y будет оставаться достаточно близким к какому-либо долгосрочному среднему значению. Чистый интегрированный процесс (случайное блуждание) в разностной форме будет иметь следующий вид:

$$\text{данные} - \text{предыдущее значение} = \delta + \text{случайный шум};$$
$$Y_t - Y_{t-1} = \delta + \varepsilon_t.$$

Если изменения или разности ряда являются результатом процесса авторегрессии и скользящего среднего (ARMA), то сам этот ряд соответствует **процессу авторегрессионного интегрированного скользящего среднего** (ARIMA). Такие процессы являются нестационарными: с течением времени подобный ряд, как правило, уходит все дальше и дальше от своего исходного состояния. Прогнозирование в случае ARIMA-модели осуществляется путем прогнозирования изменений ARMA-модели для разностей. Вследствие нестационарности такие прогнозы могут проявлять тенденцию к бесконечному нарастанию (или снижению), а границы прогнозов будут расширяться по мере распространения этих прогнозов на все более отдаленное будущее. Ниже приведена формула для ARIMA-процесса в разностной форме:

$$\text{изменение данных} = \delta + \varphi(\text{предыдущее значение}) +$$
$$+ \text{случайный шум} - \theta(\text{предыдущий случайный шум});$$
$$Y_t - Y_{t-1} = \delta + \varphi(Y_{t-1} - Y_{t-2}) + \varepsilon_t - \theta\varepsilon_{t-1}.$$

Долгосрочное среднее значение *изменения* в Y равно $\delta/(1-\varphi)$. Нельзя рассчитывать, что с течением времени Y останется достаточно близким к какому-либо долгосрочному среднему значению.

Возможно создание и более сложных ARIMA-моделей, включающих сезонное поведение поквартальных и помесячных рядов.

Основные термины

- Модель (model), математическая модель (mathematical model), или процесс (process), 745
- Прогноз (forecast), 745
- Границы прогноза (forecast limits), 745
- Анализ трендов и сезонных колебаний (trend-seasonal analysis), 755
- Тренд (trend), тенденция, 755
- Сезонный компонент (seasonal component), 755
- Циклический компонент (cyclic component), 755
- Нерегулярный компонент (irregular component), 755
- Отношение к скользящему среднему (ratio-to-moving-average), 756
- Скользящее среднее (moving average), 757
- Сезонный индекс (seasonal index), 760
- Сезонная поправка (seasonal adjustment), 763
- Экономная модель (parsimonious model), 773
- Процесс случайного шума (random noise process), 774

- Процесс авторегрессии (autoregressive (AR) process), 775
- Процесс скользящего среднего (moving-average (MA) process), 776
- Процесс авторегрессии и скользящего среднего (autoregressive moving-average (ARMA) process), 778
- Чистый интегрированный процесс, или случайные блуждания (pure integrated (I) process or random walk), 782
- Нестационарный процесс (nonstationary process), 782
- Стационарный процесс (stationary process), 782
- Процесс авторегрессионного интегрированного скользящего среднего (autoregressive integrated moving-average (ARIMA) process), 785

Контрольные вопросы

1. а) В чем отличие временного ряда от данных об одном временном срезе?

   б) Какая информация утрачивается, когда вы анализируете гистограмму, построенную для временного ряда?
2. а) Что такое "прогноз"?

   б) Что такое "границы прогноза"?

   в) Какую роль в прогнозировании играет математическая модель?

   г) Почему анализ трендов и сезонных колебаний не позволяет получить границы прогноза?
3. а) Назовите четыре базовых компонента помесячных или поквартальных временных рядов (с точки зрения подхода, основанного на трендах и сезонных колебаниях).

   б) Подробно опишите различия между циклическим и нерегулярным компонентами.
4. а) В чем отличие скользящего среднего от исходного ряда?

   б) Почему в случае анализа трендов и сезонных колебаний мы используем в скользящем среднем данные именно за целый год?

   в) Какие компоненты сохраняются в скользящем среднем? Какие уменьшаются или вообще исчезают?
5. а) Как вычисляется отношение к скользящему среднему? Какие компоненты оно представляет?

   б) Что нужно сделать, чтобы на основе отношения к скользящему среднему получить сезонный индекс? Почему это возможно?

   в) Что представляет собой сезонный индекс?

   г) Как внести сезонную поправку в значение временного ряда? Как вы интерпретируете полученный результат?
6. а) Как оценивается линейный тренд в анализе трендов и сезонных колебаний?

   б) Какой вид прогноза представляет линейный тренд?

   в) Как получить прогноз на основе линейного тренда?

   г) Какие компоненты будут представлены в этом прогнозе? Какие будут отсутствовать?
7. а) Каким образом гибкость ARIMA-процессов Бокса–Дженкинса помогает в анализе временных рядов?

   б) Что такое "экономная модель"?

   в) Как соотносится прогноз с фактическим будущим поведением оцениваемого процесса?

   г) Как соотносятся границы прогноза с фактическим будущим поведением оцениваемого процесса?
8. а) Дайте определение процесса случайного шума с точки зрения взаимосвязи между последовательными наблюдениями.

   б) Прокомментируйте следующее утверждение: если мы имеем дело с процессом случайного шума, то для его анализа не требуется применять специальные методы исследования временных рядов.

   в) Что представляют собой прогноз и границы прогноза для процесса случайного шума?
9. а) Дайте определение процесса авторегрессии первого порядка с точки зрения взаимосвязи между последовательными наблюдениями.

   б) Что представляют собой переменные X и Y в регрессионной модели для прогнозирования следующего наблюдения в процессе авторегрессии первого порядка?

   в) Опишите прогнозы процесса авторегрессии в терминах последнего наблюдения и долгосрочного среднего значения для оцениваемой модели.
10. а) Дайте определение процесса скользящего среднего в терминах взаимосвязи между последовательными наблюдениями.

    б) Какое скользящее среднее (скользящее среднее *чего именно*) мы имеем в виду, когда говорим о "процессе скользящего среднего"?

    в) Для процесса скользящего среднего первого порядка опишите в терминах долгосрочного среднего значения для оцениваемой модели прогнозы на два или больше периодов времени в будущее.
11. а) Дайте определение ARMA-процесса первого порядка в терминах взаимосвязи между последовательными наблюдениями.

    б) Значение какого параметра ARMA-процесса нужно установить равным нулю, чтобы получить процесс авторегрессии?

    в) Значение какого параметра ARMA-процесса нужно установить равным нулю, чтобы получить процесс скользящего среднего?

    г) Опишите прогнозы на отдаленное будущее исходя из ARMA-процесса.
12. а) Дайте определение случайного блуждания в терминах взаимосвязи между последовательными наблюдениями.

    б) Подробно опишите различия между процессом случайного шума и случайным блужданием.

в) Прокомментируйте следующее утверждение: если мы имеем дело со случайным блужданием, то для его анализа не требуется применять специальные методы исследования временных рядов.

г) Каково влияние составляющей дрейфа в случайном блуждании?

д) Опишите прогнозы для процесса случайного блуждания.

13. Чем различается поведение стационарных и нестационарных временных рядов?

14. Для каждого из перечисленных ниже видов процессов укажите, является ли он стационарным или нестационарным.

а) Процесс авторегрессии.

б) Случайное блуждание.

в) Процесс скользящего среднего.

г) ARMA-процесс.

15. а) Дайте определение ARIMA-процесса первого порядка в терминах взаимосвязи между последовательными наблюдениями.

б) Значение какого параметра ARIMA-процесса нужно установить равным нулю, чтобы получить случайное блуждание?

в) Как получить ARMA-процесс из ARIMA-процесса?

г) Опишите прогнозы на отдаленное будущее исходя из ARIMA-процесса.

16. Какие потребуются дополнительные члены уравнений, чтобы включить сезонное поведение в усовершенствованные ARIMA-модели?

Задачи

1. Для каждого из перечисленных ниже случаев укажите, присутствует ли в нем значительный сезонный компонент. Поясните свой ответ.

а) Продажа цветной оберточной бумаги (объемы продаж фиксируются помесячно).

б) Количество авиапассажиров, направляющихся на Гавайи из Чикаго (количество пассажиров фиксируется помесячно).

в) Биржевой индекс S&P 500 (фиксируется ежедневно). Предполагается, что биржа работает эффективно, в результате чего любые прогнозируемые тенденции уже устранены действиями крупных инвесторов, пытающихся извлечь из них для себя выгоду.

2. Некоторое время вас терзают подозрения, что проблемы с производством обостряются, как правило, именно в зимние месяцы — в первом квартале каждого года. Анализ трендов и сезонных колебаний процента производственного брака позволил установить следующие значения сезонных индексов: 1,00 — 1-й квартал; 1,01 — 2-й квартал; 1,03 — 3-й квартал и 0,97 — 4-й квартал. Подтверждает ли этот анализ ваши подозрения о том, что наивысший процент производственного брака приходится именно на первый квартал? Если да, обоснуйте свой ответ. Если нет, тогда, может быть, следует обратить внимание на какой-то другой квартал?

3. В январе у одного из банков зафиксировано 38 091 операция в сети автоматических кассовых аппаратов, а в феврале — 43 182. Соответствующий сезонный индекс для января равен 0,925, а для февраля — 0,986.

а) На какой процент увеличилось количество операций в сети автоматических кассовых аппаратов с января по февраль?

б) На какой процент должно было бы, по вашему мнению, увеличиться количество операций в сети автоматических кассовых аппаратов с января по февраль? (*Подсказка:* воспользуйтесь сезонными индексами.)

в) Определите, учитывая сезонную поправку, количество операций в сети автоматических кассовых аппаратов для каждого из этих двух месяцев.

г) На какой процент увеличилось (или уменьшилось) количество операций в сети автоматических кассовых аппаратов с января по февраль с учетом сезонной поправки?

4. На производственном собрании все выразили удовлетворение тем фактом, что объем продаж в фирме вырос с $21 791 000 в третьем квартале до $22 675 000 в четвертом квартале. Кратко опишите анализ этой ситуации (с учетом поправки на сезон), если вам известно, что сезонный индекс для третьего квартала равен 1,061, а для четвертого — 1,180. Так ли радужна картина с объемами продаж в вашей фирме, как показалось участникам собрания?

5. В табл. 14.4.1 представлены поквартальные величины нетто-продажи (суммарные продажи компании за вычетом возврата продукции, штрафов, расходов по доставке, скидок и т.п.) и доходы компании Deere & Company — крупного производителя сельскохозяйственного и промышленного оборудования.

а) Постройте график временного ряда для этой совокупности данных. Опишите все тенденции и сезонные колебания, замеченные вами на этом графике.

б) Вычислите скользящее среднее (используя каждый раз данные за один год) для этого временного ряда. Постройте график временного ряда, содержащий и данные, и скользящее среднее.

в) Найдите сезонный индекс для каждого квартала. Кажутся ли полученные вами значения обоснованными, если исходить из построенного графика временного ряда?

г) Какой из кварталов (1, 2, 3 или 4) оказывается для Deere & Company самым неблагоприятным? Насколько ниже (в среднем) оказывается объем продажи в этом квартале по сравнению с типичным кварталом в течение года?

д) Определите значения объемов продажи с поправкой на сезон, соответствующие каждому из исходных величин объемов продаж.

е) С третьего по четвертый квартал 1995 г. объемы продажи увеличились с 2 673 до 2 718. Как выглядит картина с учетом сезонной поправки?

ж) Со второго по третий квартал 1997 г. объемы продажи Deere & Company снизились с 3 521 до 3 430. Как выглядит картина с учетом сезонной поправки?

Таблица 14.4.1. Квартальные объемы продажи компании Deere & Company

Год	Нетто-продажи и доходы, млн дол.	Год	Нетто-продажи и доходы, млн дол.
1995	2 088	1996	2 905
1995	2 812	1996	2 917
1995	2 673	1997	2 396
1995	2 718	1997	3 512
1996	2 318	1997	3 430
1996	3 089	1997	3 444

Данные получены из ежегодных отчетов и краткой информации о деятельности компании.

з) Найдите уравнение регрессии для прогнозирования долгосрочного тренда изменения объемов продажи (с учетом сезонной поправки) для каждого периода времени, используя в качестве значений переменной X числа 1, 2, ...

и) Вычислите прогноз (с поправкой на сезон) на второй квартал 2000 г.

к) Вычислите прогноз на первый квартал 2001 г.

6. В табл. 14.4.2 приведены данные о квартальных объемах продажи Castle & Cooke, Inc. — международной компании, специализирующейся на производстве известных марок продуктов питания (Dole, Bumble Bee, A&W и др.). В годовом отчете за 1983 г. утверждается, что "значительное влияние на ежеквартальные результаты деятельности компании оказывают сезонные факторы, неразрывно связанные с ее бизнесом".

а) Постройте график временного ряда для этой совокупности данных. Согласны ли вы, что в этом случае действительно имеют место сезонные факторы?

б) Вычислите скользящее среднее (используя каждый раз данные за один год) для этого временного ряда. Постройте график временного ряда, включающий как данные, так и значения скользящего среднего.

в) Опишите циклическое поведение (если оно наблюдается) скользящего среднего.

г) Найдите сезонный индекс для каждого квартала. Кажутся ли полученные вами значения обоснованными, если исходить из построенного графика временного ряда?

д) Какой из кварталов (1, 2, 3 или 4) оказывается для Castle & Cooke самым благоприятным? Насколько в среднем выше объем продажи в этом квартале по сравнению с типичным кварталом в течение года?

е) Какой из кварталов (1, 2, 3 или 4) оказывается для Castle & Cooke самым неблагоприятным? Насколько в среднем ниже оказывается объем продажи в этом квартале по сравнению с типичным кварталом в течение года?

ж) Определите значения объемов продажи с учетом сезонной поправки, соответствующие каждой из исходных величин объема продажи. Постройте график для этого временного ряда с поправкой на сезон.

Таблица 14.4.2. Квартальные объемы продажи компании Castle & Cooke, Inc.

Год	Объем продажи, тыс. дол.	Год	Объем продажи, тыс. дол.
1982	453 491	1984	343 167
1982	343 669	1984	468 195
1982	387 988	1985	460 398
1982	435 642	1985	324 155
1983	352 004	1985	386 082
1983	284 030	1985	429 918
1983	320 867	1986	381 080
1983	404 634	1986	487 473
1984	402 120	1986	492 266
1984	306 606	1986	377 072

Данные получены из годовых отчетов компании Castle & Cooke, Inc. за 1983–1986 гг.

з) Опишите поведение этого временного ряда с поправкой на сезон. В частности, выявите любые изменения непостоянства продаж за этот период времени.

7. В табл. 14.4.3 приведены данные о квартальных объемах продажи компании Nordstrom, Inc. и ее филиалов.

а) Постройте график временного ряда для этой совокупности данных. Опишите сезонное и циклическое поведение, замеченное вами на этом графике. Укажите также любые свидетельства нерегулярного поведения.

б) Какой (или какие) из кварталов оказывается для Nordstrom самым благоприятным с точки зрения объемов продажи, если исходить из графика, построенного вами в п. "а"?

в) Можно ли считать сезонную картину (исходя из графика, построенного вами в п. "а") повторяющейся на протяжении всего периода времени?

г) Вычислите скользящее среднее (используя каждый раз данные за один год) для этого временного ряда. Постройте график временного ряда, включающий как значения данных, так и значения скользящего среднего.

д) Опишите циклическое поведение, выявленное с помощью скользящего среднего.

е) Найдите сезонный индекс для каждого квартала. Кажутся ли полученные вами значения обоснованными, если исходить из построенного графика временного ряда?

ж) Определите значения объемов продажи с поправкой на сезон, соответствующие каждой из исходных величин объема продажи. Постройте график для этого временного ряда с поправкой на сезон.

з) Замечаете ли вы в целом линейную долгосрочную тенденцию к увеличению или уменьшению объемов продажи компании? Можно ли использовать линию регрессии для прогнозирования этого ряда?

8. Исходя из накопленных за несколько прошлых лет данных вы обнаружили сезонные колебания объемов продажи в своей фирме. Сезонный индекс за ноябрь равен 1,08; за декабрь — 1,38 и за январь — 0,84. Объем продажи в ноябре составил $285 167.

   а) Можно ли, как правило, ожидать увеличения объемов продажи с ноября по декабрь в "типичном" году. Обоснуйте свой ответ.

   б) Найдите объем продажи в ноябре с поправкой на сезон.

   в) Внесите в показатель объема продажи в ноябре (с поправкой на сезон) сезонность с помощью декабрьского индекса, чтобы найти ожидаемый объем продажи в декабре.

   г) Вам объявили, что объем продажи в декабре составил $430 106. Оказался ли этот показатель выше или ниже, чем ожидалось, если исходить из объема продажи в ноябре?

   д) Найдите объем продажи в декабре с поправкой на сезон.

   е) Объемы продажи с ноября по декабрь — с учетом поправки на сезон — выросли или, наоборот, снизились? О чем это свидетельствует?

   ж) Пользуясь тем же методом, что и в п. "в", найдите ожидаемый объем продаж в январе исходя из объема продажи в декабре.

9. Вы решили изучить поквартальное количество посетителей своего ресторана для любителей горнолыжного спорта, воспользовавшись методом анализа трендов и сезонных колебаний. Квартальные сезонные индексы равны 1,45; 0,55; 0,72 и 1,26 для 1-го, 2-го, 3-го и 4-го кварталов соответственно. Линейный тренд оценивается уравнением вида 5 423 + 408(номер квартала), причем номер квартала начинается с 1 в первом квартале 1997 г. и увеличивается на единицу для каждого последующего квартала.

   а) Найдите прогнозируемое значение (с поправкой на сезон) для первого квартала 2001 г.

   б) Найдите прогнозируемое значение (с поправкой на сезон) для второго квартала 2001 г.

   в) Почему прогнозируемое значение с поправкой на сезон оказалось большим во втором квартале, в котором, как можно было бы предположить, ресторан посещает меньшее количество лыжников?

Таблица 14.4.3. Поквартальные нетто-продажи компании Nordstrom, Inc.

Год	Нетто-продажи, млн дол.	Год	Нетто-продажи, млн дол.
1995	816	1996	984
1995	1 149	1996	1 321
1995	907	1997	954
1995	1 242	1997	1 353
1996	906	1997	1 090
1996	1 241	1997	1 455

Данные получены из ежегодных отчетов компании.

   г) Найдите прогнозируемое значение для первого квартала 2001 г.

   д) Найдите прогнозируемое значение для второго квартала 2001 г.

   е) С учетом поправки на сезон и в соответствии с оценкой линейного тренда ответьте на вопрос: насколько больше посетителей вы ожидаете обслуживать в своем ресторане каждый квартал в сравнении с предыдущим кварталом?

   ж) Ваш стратегический бизнес-план включает проект значительного расширения ресторанного бизнеса (количество посетителей ресторана должно достичь 70 000 за год). В каком году — в соответствии с вашим прогнозом — это должно произойти впервые? (*Подсказка:* вычислите и сложите четыре прогнозируемых значения для каждого года, чтобы найти годовые итоговые показатели для 2003 и 2004 гг.)

10. Рассмотрим временной ряд квартальных объемов продажи (в тысячах), представленный в табл. 14.4.4. Квартальные сезонные индексы равняются 0,89; 0,88; 1,27 и 0,93 для 1-го, 2-го, 3-го и 4-го кварталов соответственно.

    а) Для каждого из значений объема продаж, приведенных в таблице, найдите соответствующие величины объема продаж с поправкой на сезон.

    б) В каком квартале торговля ведется наиболее активно?

    в) Как следует из приведенных данных, объем продаж в период со второго по третий квартал 1998 г. вырос с 817 до 1 073. Что можно сказать об изменениях за этот период, если учитывать поправку на сезон?

    г) Как следует из приведенных данных, объем продаж в период с третьего по четвертый квартал 1997 г. снизился с 1 084 до 819. Что можно сказать об изменениях за этот период, если учитывать поправку на сезон?

    д) Значения экспоненциального тренда для четырех кварталов 2002 г. равняются 1964, 2070, 2183 и 2301. Внесите поправку на сезон в эти прогнозы тренда, чтобы получить прогнозы фактических продаж для 2002 г.

11. Какой тип анализа временных рядов обеспечит получение простейших результатов для изучения спроса на мазут (используемый для обогрева), который, как правило, достигает пика в зимний период?

12. Ваш прогноз помесячных объемов продажи (с поправкой на сезон) выражается формулой $382 190 + $4 011(номер месяца), причем номер месяца равен 1 для января и затем последовательно увеличивается на 1 для каждого следующего месяца. Сезонный индекс для февральских продаж равен 0,923; для апреля этот индекс равен 1,137. Все, что вам сейчас требуется, — это прогноз стоимости реализованных товаров, чтобы заранее спланировать выполнение заказов. Вы выяснили, что месячные объемы продажи являются хорошим показателем (прогнозом) стоимости продукции, реализованной за месяц, и составили следующее уравнение регрессии:

    прогнозируемая стоимость реализованных товаров =
    = $106 582 + 0,413(объем продажи).

    а) Найдите прогнозируемое значение месячного объема продаж (с поправкой на сезон) в феврале 2002 г.

б) Найдите прогнозируемое значение месячного объема продаж в феврале 2002 г.

в) Найдите прогнозируемое значение стоимости товаров, проданных в феврале 2002 г.

г) Найдите прогнозируемое значение стоимости товаров, проданных в апреле 2003 г.

13. Для каждой из перечисленных ниже ситуаций укажите, какому типу процесса (стационарный или нестационарный) она соответствует.

а) Цена одной акции компании IBM, фиксируемая ежедневно.

б) Прайм-рейт, фиксируемый еженедельно и представляющий собой публикуемую банками процентную ставку по кредитам для наилучших заемщиков.

в) Толщина бумаги, измеряемая пять раз в минуту в процессе производства бумаги и ее намотки на рулоны. (Предполагается, что этот процесс находится под контролем.)

г) Цена одной страницы рекламного объявления в журнале *TV Guide*; изменяется раз в год.

14. В табл. 14.4.5 представлены основные результаты компьютерного анализа методом Бокса–Дженкинса ежедневных процентных изменений промышленного

Таблица 14.4.4. Квартальные объемы продажи

Квартал	Год	Объем продаж, тыс. дол.	Квартал	Год	Объем продаж, тыс. дол.
1	1995	438	1	1997	676
2	1995	432	2	1997	645
3	1995	591	3	1997	1 084
4	1995	475	4	1997	819
1	1996	459	1	1998	710
2	1996	506	2	1998	817
3	1996	736	3	1998	1 073
4	1996	542			

Таблица 14.4.5. Результаты анализа методом Бокса–Дженкинса ежедневных процентных изменений индекса Доу Джонса

Коэффициент	Оценка	Стандартная ошибка	*t*-отношение
Авторегрессия	–0,3724	1,7599	–0,21
Скользящее среднее	–0,4419	1,6991	–0,26
Константа	–0,000925	0,002470	–0,37
Среднее значение	–0,000674	0,001799	–0,37
Стандартное отклонение случайного шума	0,011950		

индекса Доу Джонса с 31 июля по 9 октября 1987 г. (до наступления известного краха 1987 г.).

а) Процесс какого типа мы оцениваем в данном случае?

б) Запишите модель таким образом, чтобы из нее было видно, как следующее наблюдение определяется на основе предыдущего. Воспользуйтесь фактическими оценками коэффициентов.

в) Какие оценки коэффициентов являются статистически значимыми?

г) Подставив 0 вместо всех тех оценок коэффициентов, которые не являются статистически значимыми, запишите модель, которая показывала бы, как следующее наблюдение определяется на основе предыдущего. Процесс какого типа мы получили в этом случае?

д) В одном абзаце опишите полученные результаты в поддержку применимости теории случайных блужданий к поведению фондовой биржи.

15. Объявления о срочном найме используются компаниями в тех случаях, когда они попадают в "цейтнот" и им срочно требуются работники (это бывает, например, когда компания стремительно наращивает свою активность после преодоления периода спада). В табл. 14.4.6 представлены результаты анализа методом Бокса–Дженкинса индекса рекламных объявлений о срочном найме, а на рис. 14.4.1 показан соответствующий временной ряд с прогнозами, полученными методом Бокса–Дженкинса.[21]

а) Процесс какого типа мы оцениваем в данном случае?

б) Какие оценки коэффициентов являются статистически значимыми (если таковые действительно имеются)?

в) Имея в своем распоряжении график, показанный на рис. 14.4.1, были бы вы удивлены, если бы индекс рекламных объявлений о срочном найме упал в 1995 г. до значения 15?

г) Имея в своем распоряжении график, показанный на рис. 14.4.1, были бы вы удивлены, если бы индекс рекламных объявлений о срочном найме вырос в 1996 г. до значения 120?

д) Исходя из рис. 14.4.1 создается впечатление, что примерно после 1995 г. прогнозируемые значения выравниваются. Свидетельствует ли это о том, что индекс рекламных объявлений о срочном найме в будущем уже не будет меняться год от года? Поясните свой ответ.

16. В табл. 14.4.7 и 14.4.8 представлены основные результаты компьютерного анализа методом Бокса–Дженкинса доходов по казначейским векселям США за год в период с 1961 по 1990 гг.

а) Процесс какого типа мы оцениваем в данном случае?

б) Запишите модель таким образом, чтобы из нее было видно, как следующее наблюдение определяется на основе предыдущего.

в) Какие оценки коэффициентов являются статистически значимыми?

[21] Данные получены из *Business Statistics 1963–1991* (Washington D.C.: U.S. Government Printing Office, June 1992), p. 59.

Таблица 14.4.6. Результаты анализа методом Бокса–Дженкинса индекса рекламных объявлений о срочном найме

Окончательные оценки параметров

Тип	Оценка	Стандартное отклонение	t-отношение
AR 1	0,5331	0,1780	2,99
MA 1	–0,8394	0,1170	–7,18
Константа	50,653	5,302	9,55
Среднее значение	108,48	11,36	

Количество наблюдений: 29

Остатки: SS = 6161,77 (прогнозы относительно прошлого исключены)
MS = 236,99 DF = 26

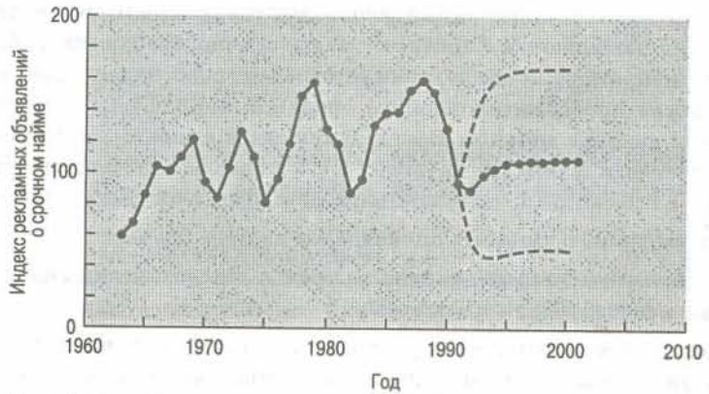


Рис. 14.4.1. Изменение величины индекса рекламных объявлений о срочном найме с 1963 по 1991 гг.; указаны также прогнозы и 95% интервалы на 10 лет вперед (на основе модели временных рядов Бокса–Дженкинса)

г) Начертите график временного ряда для исходных данных (на основе табл. 14.1.5), прогнозы и границы прогнозов.

д) Прокомментируйте эти прогнозы и границы прогнозов.

17. Рабочая неделя на протяжении многих лет постепенно — в среднем — сокращается. В табл. 14.4.9 представлены результаты компьютерного анализа методом Бокса–Дженкинса средней продолжительности рабочей недели (в часах), а на рис. 14.4.2 показан соответствующий временной ряд с прогнозами, выполненными методом Бокса–Дженкинса.[22]

[22] Данные получены из Министерства экономики США (Бюро экономического анализа), *Business Statistics 1963–1991* (Washington D.C.: U.S. Government Printing Office, June 1992), p. 50. Они представляют собой среднее количество рабочих часов в неделю на одного производственного или не принадлежащего к руководящему составу работника, занятого на частных несельскохозяйственных предприятиях, без поправки на сезон. Интересно отметить, что среднее количество рабочих часов в неделю исключительно для заводских рабочих не имеет такой тенденции к сокращению на протяжении ряда лет.

Таблица 14.4.7. Результаты анализа методом Бокса–Дженкинса процентных ставок по казначейским векселям США

Коэффициент	Оценка	Стандартная ошибка	*t*-отношение
Авторегрессия	0,644	0,156	4,13
Скользящее среднее	–0,755	0,134	–5,62
Константа	2,199	0,446	4,93
Среднее значение	6,168	1,250	4,93
Стандартное отклонение случайного шума	1,385		

Таблица 14.4.8. Результирующие прогнозы, полученные из анализа методом Бокса–Дженкинса процентных ставок по казначейским векселям США

		95% границы прогноза	
Год	Прогноз	нижняя	верхняя
1991	5,806	3,091	8,521
1992	5,935	1,267	10,602
1993	6,018	0,750	11,286
1994	6,071	0,574	11,569
1995	6,106	0,516	11,696
1996	6,128	0,500	11,755
1997	6,142	0,499	11,785
1998	6,151	0,501	11,801
1999	6,157	0,505	11,809
2000	6,161	0,507	11,814

Таблица 14.4.9. Результаты анализа методом Бокса–Дженкинса средней продолжительности рабочей недели

Окончательные оценки параметров

Тип	Оценка	Стандартное отклонение	t-отношение
AR 1	0,0086	0,1963	0,04
Константа	–0,15943	0,03934	–4,05

Вычисление разностей: 1 обычная разность

Количество наблюдений: исходный ряд – 29; после вычисления разностей – 28

Остаточные значения: SS = 1,12671 (прогнозы относительно прошлого исключены)
MS = 0,04334 DF = 26

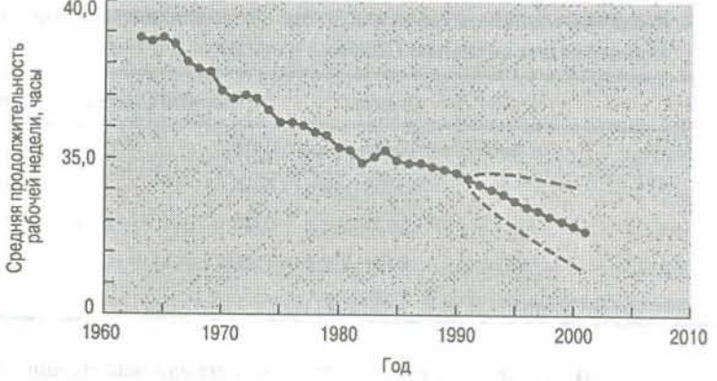


Рис. 14.4.2. Средняя продолжительность рабочей недели (в часах) в США за период с 1963 по 1991 гг.; указаны также прогнозы и 95% интервалы на 10 лет вперед (на основе модели временных рядов Бокса–Дженкинса)

а) Какой тип компонента (авторегрессия или скользящее среднее) включает эта оценочная модель?

б) Является ли компонент модели, названный вами в п. "а", статистически значимым?

в) Является ли статистически значимым постоянный член (константа)? Интерпретируйте оценку константы как ежегодный темп сокращения средней продолжительности рабочей недели.

г) Если исходить из рис. 14.4.2, то насколько неожиданным для вас был бы тот факт, что средняя продолжительность рабочей недели в 1995 г. равнялась 33,5 часа?

д) Если исходить из рис. 14.4.2, то насколько неожиданным для вас был бы тот факт, что средняя продолжительность рабочей недели в 2000 г. равнялась 37,5 часа?

Проекты

1. Выберите какую-либо интересующую вас фирму и получите данные о поквартальных объемах продаж этой фирмы по крайней мере за три последовательных года (для этого можно воспользоваться ежегодными отчетами фирмы, которые можно получить в библиотеке или через Internet).

   а) Изобразите график временного ряда и прокомментируйте структуру, которая следует из этого графика.

   б) Вычислите скользящее среднее за год, отобразите его на своем графике и прокомментируйте.

   в) Вычислите сезонные индексы, отобразите их на своем графике и прокомментируйте.

   г) Вычислите и отобразите на своем графике временной ряд с поправкой на сезон, затем прокомментируйте полученный результат. В частности, ответьте на вопрос: какую новую информацию можно извлечь в результате внесения сезонной поправки?

   д) Вычислите линию тренда и внесите в нее сезонную поправку, чтобы получить прогнозы на два последующих года. Отобразите эти прогнозы на своем графике наряду с исходными данными. Прокомментируйте, насколько правдоподобными кажутся вам эти прогнозы.

2. 
   Подберите ежегодные (как минимум за 20 последовательных лет — в виде временного ряда) данные экономического характера, представляющие для вас интерес. (Реализация этого проекта связана с доступом к компьютерному программному обеспечению, которое позволяет оценивать ARIMA-модели.)

   а) Представьте этот временной ряд в графическом виде и прокомментируйте структуру, которая следует из полученного графика.

   б) Процессу какого типа соответствует исследуемый ряд — стационарному или нестационарному? Если он относится к числу крайне нестационарных (например, заканчивается далеко от того значения, с которого начинался), отобразите в графическом виде разности, чтобы выяснить, являются ли они стационарным процессом.

   в) Примените к своему ряду (или к соответствующим разностям, если ряд оказался нестационарным) процесс авторегрессии первого порядка. Является ли коэффициент авторегрессии статистически значимым, если исходить из t-статистики?

   г) Оцените соответствие своего ряда (или соответствующих разностей, если ряд оказался нестационарным) процессу скользящего среднего первого порядка. Является ли коэффициент скользящего среднего статистически значимым, если исходить из t-статистики?

   д) Оцените соответствие своего ряда (или соответствующих разностей, если ряд оказался нестационарным) ARMA-процессу первого порядка. Какие коэффициенты являются статистически значимыми, если исходить из t-статистики?

   е) Основываясь на результатах применения трех перечисленных моделей, укажите ту модель, которой вы отдали бы предпочтение. (Можно исключить те компоненты, которые не являются статистически значимыми.)

   ж) Теперь можно вернуться к исходному ряду (даже если вы пользовались разностями). Оцените выбранную вами модель — включив интегрированный (И) компонент, если вы пользовались разностями, — и определите прогнозы и границы прогнозов.

   з) Представьте полученные вами прогнозы и границы прогнозов в графическом виде наряду с исходными данными и прокомментируйте результат.

   и) Прокомментируйте процедуру выбора модели. (Помните, что процедура выбора модели намного усложняется при использовании процессов более высокого порядка.)

Методы и применения

В этой части...

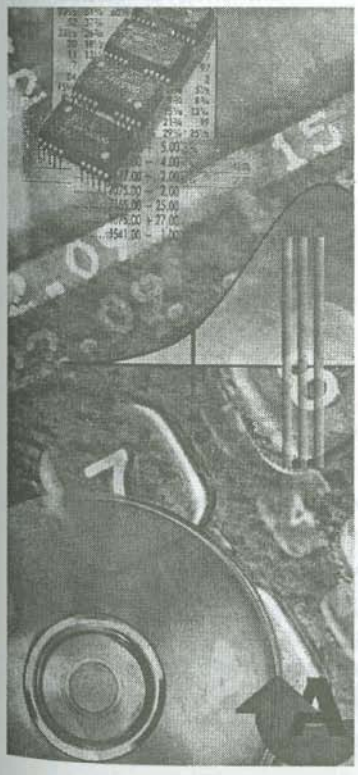
Глава 15. "Дисперсионный анализ: проверка различий для нескольких выборок и многое другое"

Глава 16. "Непараметрические методы: проверка гипотез для порядковых данных или данных, не подчиняющихся нормальному распределению"

Глава 17. "Анализ "хи-квадрат": поиск закономерностей для качественных данных"

Глава 18. "Контроль качества: выявление вариации и управление ею"

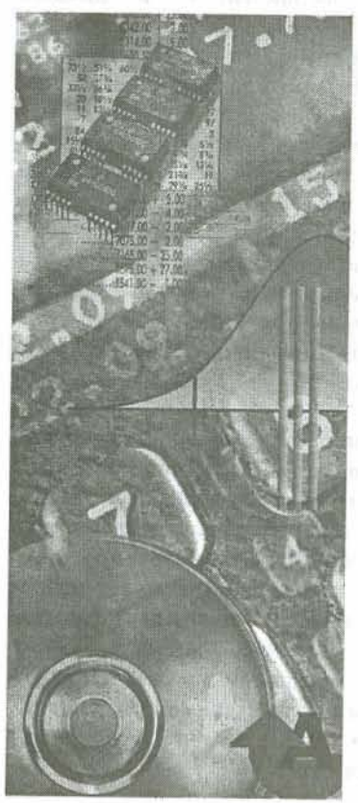
Эти последние четыре главы посвящены вопросам применения идей и методов статистики в конкретных ситуациях. *Дисперсионный анализ* (Analysis of variance, или сокращенно ANOVA) предназначен для проверок гипотез в сложных случаях. Он представляет собой группу методов сравнения размеров вариации, обусловленных различными факторами. В главе 15 вы познакомитесь с этими методами и узнаете, как применять дисперсионный анализ для проверки гипотезы о том, что несколько выборок извлечено из одной и той же генеральной совокупности. В главе 16 описаны *непараметрические* статистические методы (работающие с рангами), с помощью которых гипотезы можно проверить в некоторых сложных ситуациях, например, если распределение отлично от нормального или данные являются *порядковыми* (упорядоченными категориями, *ординальными* данными), а не количественными (числами, имеющими содержательное значение). Для *номинальных* данных (неупорядоченных категорий) необходимы специальные методы проверки гипотез, а именно *хи-квадрат* анализ (глава 17), поскольку с этими категориями нельзя выполнять арифметические действия или упорядочение (ранжирование). Наконец, глава 18 охватывает основные статистические методы контроля качества, благодаря которым вы сможете выяснить, какие проблемы подлежат разрешению, как управлять неоднородностью (вариацией) параметров продукции и определить, где нужно вмешиваться, а где можно оставить все как есть.



ГЛАВА 15

Дисперсионный анализ: проверка различий для нескольких выборок и многое другое

Дисперсионный анализ (сокращенно **ANOVA**) дает общую схему проверки статистических гипотез, основанную на тщательном изучении различных источников вариации (изменчивости, неоднородности) в сложной ситуации. Ниже приведены некоторые примеры ситуаций, в которых следует использовать дисперсионный анализ.



Ситуация первая. С целью сокращения расходов вы проанализировали пять добавок, которые, предположительно, увеличивают объем продукции химического производства. Вы десять раз запускали процесс производства с использованием каждой из добавок и десять раз без добавки. Это пример *однофакторной модели*, так как имеется один фактор ("добавка"), который имеет несколько значений (уровней). Результат представляет собой набор данных, состоящий из 6 списков значений объема продукции. Из-за обычной изменчивости, свойственной процессу, трудно объяснить, обусловлено ли любое улучшение просто случайной удачей, или данная добавка действительно лучше других. Поэтому необходимо проверить нулевую гипотезу: действительно ли эти шесть списков значений объемов продукции одинаковы и все различия между ними являются результатом только лишь случайности. Использовать для проверки t-тест, описанный в главе 10 нельзя, поскольку речь идет более чем о двух выборках.[1] Вместо такого теста применяют однофакторный дисперсионный анализ, призванный выяснить, есть ли какие-либо значимые (т.е. систематические, или неслучайные) различия между этими добавками. Если значимые различия имеются, то можно продолжить их подробное изучение. В противном случае можно сделать вывод, что фиксируемых систематических различий между добавками нет.

Ситуация вторая. В принципе, можно использовать в производственном процессе *комбинацию* добавок. При наличии 5 добавок возможно $2^5 = 32$ комбинации (включая и отсутствие добавок), причем для каждой комбинации технологический процесс необходимо запустить дважды.[2] Этот пример представляет собой *факторный план* с пятью факторами (добавками), каждая из которых изучается на двух уровнях (либо используется, либо не используется). Дисперсионный анализ такого набора данных покажет, (а) влияет ли каждая добавка на объем продукции и (б) существует ли взаимодействие влияний добавок в комбинации.

Ситуация третья. Как часть маркетингового исследования, вы проверили три вида средств массовой информации (газету, радио и телевидение) в сочетании с двумя видами рекламного объявления (прямой и косвенный методы). Каждому испытуемому в этом исследовании была продемонстрирована одна комбинация, после чего был подсчитан результат (в баллах), отражающий эффективность рекламы. Это пример двухфакторного плана (факторами являются "вид средства массовой информации" и "рекламное объявление"). Дисперсионный анализ покажет, (а) существенно ли различаются виды средств массовой информации по эффективности и (б) существует ли взаимосвязь между видом средства массовой информации и типом рекламного объявления.

Для проверки каждой гипотезы в дисперсионном анализе используют **F-тест**, основанный на **F-статистике**, которая представляет собой отношение двух дисперсий.[3] Числитель представляет собой вариацию, обусловленную конкретным интересующим нас эффектом, который мы и проверяем, а знаменатель — вариацию, обусловленную случайностью. Если это отношение больше табличного F-значения, то фактор имеет значимое влияние.

Однофакторный дисперсионный анализ, в частности, используют для проверки значимости различия средних значений нескольких различных выборок. Это самый простой вид дисперсионного анализа. Хотя более сложные ситуации

---

[1] Вы можете использовать t-тест для независимых выборок, чтобы сравнивать добавки попарно. Однако ввиду необходимости 15 таких проверок эта *группа* тестов не является обоснованной, поскольку мы не управляем вероятностью ошибки для группы. В частности, если принять истинность нулевой гипотезы о том, что различия в объемах продукции нет, то вероятность сделать неверное утверждение о том, что *некоторые пары* объемов продукции значимо различаются, может быть намного выше, чем ошибка в 5%, которую применяют для каждого *отдельного* теста. Использование F-теста позволяет удержать ошибку на уровне 5%, и затем, если F-тест оказывается значимым, можно использовать модифицированные t-тесты.

[2] Имеет смысл использовать каждую комбинацию более одного раза, если есть такая возможность, потому что это позволяет получить больше информации об изменчивости в каждой ситуации.

[3] Вспомним, что дисперсия представляет собой квадрат стандартного отклонения. Именно этот показатель изменчивости используют в дисперсионном анализе. Если бы статистика развивалась несколько иначе, возможно, мы бы сейчас сравнивали отношения стандартных отклонений с таблицей, содержащей квадратные корни из значений ныне принятой F-таблицы. Однако традиционно используют метод, основанный на дисперсии, поэтому мы будем поступать так же.

требуют более сложных вычислений, общий подход остается тем же: проверить значимость, сравнив один источник вариации (проверяемый фактор) с другим источником вариации (лежащим в основе случайности).

15.1. Использование блочных диаграмм для одновременного представления нескольких выборок

Поскольку целью дисперсионного анализа является только проверка гипотезы, предварительный анализ данных полностью зависит от вас. Следует изучить статистические характеристики (например, среднее и стандартное отклонение) и гистограммы или блочные диаграммы для каждого перечня чисел из вашего набора данных. Дисперсионный анализ позволит установить, имеются ли значимые расхождения, но, чтобы фактически увидеть оценки этих расхождений, необходимо изучить обычные статистические характеристики.

Блочные диаграммы особенно хорошо подходят для сравнения нескольких распределений, поскольку опускают несущественные детали и позволяют сконцентрировать внимание на главном. Ниже приведен порядок проверки данных, рекомендуемый при использовании блочных диаграмм или гистограмм для сравнения аналогичных показателей в ряде ситуаций.

1. Как вы считаете, насколько разумными и обоснованными выглядят данные, представленные на блочной диаграмме? Желательно выделить основные проблемы *до того*, как тратить время на работу с этими данными. Например, вы можете обнаружить, что используете неверные данные. (Может, *числа выглядят как слишком большие или слишком маленькие? Может, это* данные не за последний год?) Можно также выделить основные резко отклоняющиеся значения (выбросы), которые необходимо изучить особо, и если они представляют собой ошибки, выполнить коррекцию.
2. Отличаются ли между собой центры (медианы) блоков в диаграмме? Диаграммы дают первичную, неформальную оценку, и только дисперсионный анализ позволяет получить точный, формальный ответ. Кроме того, демонстрируют ли центры блоков какую-либо интересную особенность?
3. Постоянна ли (в разумных пределах) вариация, отраженная в разных блоках диаграммы? Это важно, поскольку дисперсионный анализ предполагает, что в генеральной совокупности эти вариации равны. Если, например, расположенные на диаграмме выше блоки (с большими значениями медианы) систематически шире (т.е. демонстрируют большую вариацию), то дисперсионный анализ может дать неверный ответ.[4]

[4] Эту проблему неравенства вариации часто можно решить путем преобразования исходных данных. Можно, например, использовать логарифм при условии, что все значения больше нуля. Далее изучают блочную диаграмму для преобразованных значений, чтобы увидеть, действительно ли проблема решена. Если в результате дисперсионного анализа будут выявлены значимые различия для логарифмической шкалы, то можно сделать вывод, что исходные группы также значимо различаются. Таким образом, для преобразованных данных интерпретация результатов дисперсионного анализа практически не меняется, но при условии применения ко всем данным одинаковых преобразований.

Пример. *Сравнение качества продукции ваших поставщиков*

В настоящее время ваша фирма закупает одинаковые электронные компоненты у трех разных поставщиков, и это вас беспокоит. Несмотря на утверждения некоторых сотрудников, что такая организация снабжения позволяет фирме устанавливать хорошие цены и обеспечивает короткое время доставки, вы обеспокоены тем фактом, что конструкция изделия должна позволять использовать для установки *наихудшую* комбинацию компонентов. Главная цель фирмы — цена и доходы. *В частности, будет ли лучше, если* фирма заключит эксклюзивный контракт только с одним поставщиком, чтобы быстро получать высококачественные компоненты по повышенной цене? В качестве подготовительной информации по этому вопросу вы рассматриваете качество компонентов, полученных от каждого из ваших поставщиков.

Только что получен набор данных. Вы просите отдел технического контроля проверить по 20 компонентов *от* каждого поставщика, отобранных случайным образом из последних поставок. Отдел фактически проверил по 21 компоненту каждого поставщика, но не все результаты измерений достаточно надежны. Оценка качества, основанная на нескольких различных параметрах, измеряется по шкале от 0 до 100 и показывает степень соответствия технических спецификаций компонента требованиям фирмы. Чем выше оценка, тем лучше. Оценка 75 и выше достаточна, чтобы компонент можно было применять для разных целей. В табл. 15.1.1 приведен набор данных и основные статистические характеристики.

В среднем компания-поставщик Consolidated имеет *самую высокую оценку качества (87,7), затем идет* компания Amalgamated (82,1) и, наконец, Bipolar (80,7). Блочная диаграмма на рис. 15.1.1 показывает, что качество Consolidated в целом выше качества двух других поставщиков, хотя есть значительное перекрытие и компоненты высшего качества поступают также от Amalgamated (с оценкой 97). В этой блочной диаграмме содержится и много дополнительной информации: нет компонентов идеального качества (с оценкой 100), разные поставщики имеют схожую вариацию качества компонентов (на что указывают размеры блоков).

Хотя компания Consolidated имеет наивысшее качество продукции, вы задались вопросом, обусловлено ли это случайным отбором этих конкретных компонентов. Останется ли этот поставщик по-прежнему в среднем наилучшим, если изучить выборки большего размера от всех поставщиков? Дисперсионный анализ дает ответ на этот вопрос, не требуя больших затрат на получение дополнительных данных.

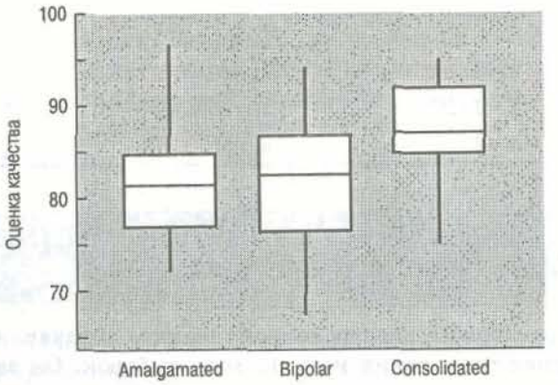


Рис. 15.1. Блочная диаграмма качества компонентов, закупаемых у трех поставщиков. Продукция компаний Amalgamated и Bipolar схожа по качеству (хотя нижние значения хуже у Bipolar). Продукция компании Consolidated постоянно характеризуется наивысшим качеством, хотя имеет место значительное перекрытие в распределении этих оценок качества

Таблица 15.1.1. Оценки качества продукции поставщиков

	Amalgamated	Bipolar	Consolidated
	75	94	90
	72	87	86
	87	80	92
	77	86	75
	84	80	79
	82	67	94
	84	86	95
	81	82	85
	78	86	86
	97	82	92
	85	72	92
	81	77	85
	95	87	87
	81	68	86
	72	80	92
	89	76	85
	84	68	93
	73	86	89
		74	83
		86	
		90	
Среднее	$\bar{X}_1 = 82{,}055556$	$\bar{X}_2 = 80{,}666667$	$\bar{X}_3 = 87{,}684211$
Стандартное отклонение	$S_1 = 7{,}124706$	$S_2 = 7{,}598245$	$S_3 = 5{,}228688$
Размер выборки	$n_1 = 8$	$n_2 = 21$	$n_3 = 19$

15.2. F-тест определяет, значимо ли различаются средние

F-тест в однофакторном дисперсионном анализе устанавливает, значимо ли различаются средние нескольких независимых выборок. Он заменяет t-тест для независимых выборок (см. главу 10) при наличии более двух выборок и дает тот же результат в случае двух выборок.

Данные и источники вариации

Набор данных в однофакторном дисперсионном анализе состоит из k независимых одномерных выборок, элементы которых измерены в одинаковых единицах (например, долларах или милях на галлон). Допустимы различные размеры выборок. Форма данных показана в табл. 15.2.1.

Таблица 15.2.1. Данные для однофакторного дисперсионного анализа

	Выборка 1	Выборка 2	...	Выборка k
	$X_{1,1}$	$X_{2,1}$	...	$X_{k,1}$
	$X_{1,2}$	$X_{2,2}$	...	$X_{k,2}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	X_{1,n_1}	X_{2,n_2}	.	X_{k,n_k}
Среднее	$\bar{X}_1$	$\bar{X}_2$	.	$\bar{X}_k$
Стандартное отклонение	S_1	S_2	.	S_k
Размер выборки	n_1	n_2	.	n_k

Чтобы определить источники вариации, следует ответить на вопрос: "Почему значения данных отличаются друг от друга?". Поскольку здесь два источника вариации, значит, и два ответа на этот вопрос.

1. Одним из источников вариации является факт отличия друг от друга генеральных совокупностей. Например, если выборка 2 включает особенно тщательную обработку, то значения в выборке 2 будут отличаться (будут выше) от значений в других выборках. Этот источник называется *межгрупповая (межвыборочная) вариация*. Чем больше межгрупповая вариация, тем очевиднее, что генеральные совокупности различаются между собой.
2. Другим источником вариации является (обычно) неоднородность значений внутри каждой выборки. Например, вы не можете ожидать, что все значения в выборке 2 будут одинаковыми. Этот источник называют *внутригрупповой (внутривыборочной) вариацией*. Чем больше внутригрупповая вариация, тем случайнее ваши данные и тем труднее установить, действительно ли различаются генеральные совокупности.

Источники вариации в однофакторном дисперсионном анализе

Межгрупповая вариация (между выборками).

Внутригрупповая вариация (внутри каждой выборки).

В примере с поставщиками двумя источниками вариации являются (1) возможно различные уровни качества трех поставщиков и (2) возможно различные оценки качества различных компонентов, полученных от *одного поставщика*.

F-тест будет основан на отношении величин этих источников вариации. Но сначала рассмотрим основные положения проверки этой гипотезы.

Допущения

Допущения, лежащие в основе F-теста в однофакторном дисперсионном анализе, обеспечивают прочный фундамент для формулировки точного вероятностного утверждения, основанного на наблюдаемых данных.

Допущения для однофакторного дисперсионного анализа

1. Набор данных состоит из k случайных выборок из k генеральных совокупностей.

2. Все генеральные совокупности имеют нормальное распределение и одинаковые стандартные отклонения, т.е. $\sigma_1 = \sigma_2 = \ldots = \sigma_k$. Это позволяет использовать для проверки гипотезы стандартные статистические таблицы.

Обратите внимание, что здесь отсутствует допущение о средних значениях нормальных распределений. Средние могут принимать любые значения — процедура проверки гипотез будет работать для любых значений средних.

В примере с поставщиками и качеством допущения заключаются в том, что данные отдела контроля качества представляют собой три случайные выборки, каждая из генеральной совокупности оценок качества продукции одного из поставщиков. Более того, предполагается, что для каждого поставщика распределение оценок является нормальным и все три поставщика имеют одинаковую вариацию оценок качества (стандартное отклонение генеральной совокупности). Проведенный ранее анализ блочной диаграммы свидетельствует о том, что допущения относительно нормальных распределений и равенства вариации разумны и приблизительно удовлетворяются для этого набора данных.

Гипотезы

Нулевая гипотеза для F-теста в однофакторном дисперсионном анализе утверждает, что k генеральных совокупностей (представленных k выборками) имеют одно и то же среднее значение. Альтернативная гипотеза утверждает, что средние *не* все равны между собой, т.е. по меньшей мере, у двух совокупностей средние различаются.

Гипотезы для однофакторного дисперсионного анализа

Нулевая гипотеза:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k \text{ (все средние равны между собой).}$$

Альтернативная гипотеза:

$$H_1: \mu_i \neq \mu_j \text{ по крайней мере для одной пары генеральных совокупностей (не все средние равны).}$$

Поскольку предполагается, что стандартные отклонения всех совокупностей равны между собой, нулевая гипотеза фактически сводится к тому, что *генеральные совокупности идентичны* (с точки зрения распределения). Альтернативная гипотеза утверждает, что существуют различия без учета количества фактически отличающихся между собой генеральных совокупностей. Иными словами, альтернативная гипотеза включает случаи, когда только одна совокупность отличается от остальных, несколько совокупностей различны между собой, и все совокупности различаются между собой.

В примере с качеством продукции поставщиков нулевая гипотеза утверждает, что все три поставщика имеют одинаковые характеристики качества: оценки качества производимых ими компонентов имеют одинаковые распределения (одно и то же нормальное распределение с одними и теми же средним значением и стандартным отклонением). Альтернативная гипотеза утверждает, что некоторые

поставщики различаются с точки зрения среднего уровня качества компонента (могут различаться все три поставщика, или качество двух может быть одинаковым, а третьего — выше или ниже).

F-статистика

F-статистика для однофакторного дисперсионного анализа представляет собой отношение значений вариации из двух источников: межгрупповую вариацию делят на внутригрупповую. Можно считать, что F-статистика показывает, во сколько раз выборочные средние более изменчивы по сравнению с тем, что можно было бы ожидать, если бы расхождение было случайным. Для выполнения F-теста необходимо вычислить F-статистику и сравнить полученное значение с табличным. Это требует некоторых дополнительных вычислений.

Поскольку нулевая гипотеза утверждает, что средние всех генеральных совокупностей равны, необходимо оценить это среднее значение, которое объединяет всю информацию о выборках. **Общее (главное) среднее** представляет собой среднее всех значений из всех выборок. Его можно также рассматривать как *взвешенное среднее* выборочных средних, где большие выборки имеют больший вес.

Общий объем выборки, n, и общее среднее, $\bar{X}$

$$n = n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i;$$

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n} = \frac{\sum_{i=1}^{k} n_i\bar{X}_i}{n}.$$

В примере о качестве продукции поставщиков общий объем выборки и общее среднее вычисляются следующим образом:

$$n = n_1 + n_2 + \ldots + n_k = 18 + 21 + 19 = 58;$$

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n} =$$

$$= \frac{(18 \times 82{,}055556) + (21 \times 80{,}666667) + (19 \times 87{,}684211)}{58} =$$

$$= \frac{4837}{58} = 83{,}396552$$

Межгрупповая вариация показывает, насколько различаются выборочные средние. Она равна нулю, если средние равны, и тем больше, чем сильнее различаются средние. Межгрупповая вариация представляет основу меры вариации выборочных средних.[5] Соответствующая формула приведена ниже.

[5] Формулу для вычисления межгрупповой вариации можно рассматривать как результат замены всех значений во всех выборках соответствующими выборочными средними, объединения результатов этих замен в один большой набор данных, вычисления обычного выборочного стандартного отклонения, возведения его в квадрат для получения дисперсии и умножения на масштабирующий коэффициент $(n-1)/(k-1)$.

Межгрупповая вариация для однофакторного дисперсионного анализа

$$\text{Межгрупповая вариация} = \frac{n_1(\bar{X}_1-\bar{X})^2+n_2(\bar{X}_2-\bar{X})^2+\ldots+n_k(\bar{X}_k-\bar{X})^2}{k-1} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})^2}{k-1}.$$

Число степеней свободы = $k-1$.

Число степеней свободы отражает тот факт, что измеряют вариацию k средних. Одну степень свободы при этом вычитают (как и при обычном стандартном отклонении), поскольку для общего среднего оценка известна.

В примере с качеством продукции поставщиков межгрупповая вариация для числа степеней свободы, равного $k-1=3-1=2$, вычисляется следующим образом:

$$\text{межгрупповая вариация} = \frac{n_1(\bar{X}_1-\bar{X})^2+n_2(\bar{X}_2-\bar{X})^2+\ldots+n_k(\bar{X}_k-\bar{X})^2}{k-1} =$$

$$= \begin{bmatrix} 18(82,055556-83,39552)^2+21(80,666667-83,396552)^2 \\ +19(87,684211-83,396552)^2 \end{bmatrix} / (3-1) =$$

$$= \frac{32,369+156,498+349,296}{2} = \frac{538,163}{2} = 269,08$$

Внутригрупповая вариация измеряет, насколько неоднородна каждая выборка. Все выборки имеют одинаковую вариацию, т.е. только одно значение внутригрупповой вариации. Это значение должно быть равно нулю, если каждая выборка состоит из своих многократно повторенных средних, и оно будет тем больше, чем сильнее различаются числа в выборке. Квадратный корень из значения внутригрупповой вариации дает оценку стандартных отклонений для генеральных совокупностей. Соответствующая формула приведена ниже.

Внутригрупповая вариация для однофакторного дисперсионного анализа

$$\text{Внутригрупповая вариация} = \frac{(n_1-1)(S_1)^2+(n_2-1)(S_2)^2+\ldots+(n_k-1)(S_k)^2}{n-k} = \frac{\sum_{i=1}^{k}(n_i-1)(S_i)^2}{n-k}$$

Число степеней свободы = $n-k$.

Число степеней свободы отражает тот факт, что измеряют вариацию всех n значений данных относительно их выборочных средних, но при этом вычитают k степеней свободы, поскольку известны оценки k различных выборочных средних.

В примере о качестве продукции поставщиков внутригрупповую вариацию (с $n-k=58-3=55$ степеней свободы) вычисляют следующим образом:

$$\text{внутригрупповая вариация} = \frac{(n_1-1)(S_1)^2+(n_2-1)(S_2)^2+\ldots+(n_k-1)(S_k)^2}{n-k} =$$

$$= \frac{(18-1)(7,124706)^2+(21-1)(7,598245)^2+(19-1)(5,228688)^2}{58-3} =$$

$$= \frac{(17\times 50,7614)+(20\times 57,7333)+(18\times 27,3392)}{55} =$$

$$= \frac{862,944+1154,667+192,105}{55} = \frac{2509,716}{55} = 45,63$$

F-статистика представляет собой отношение этих двух значений вариации (межгрупповой и внутригрупповой), показывая, в какой мере выборочные средние различаются между собой (числитель) с учетом общего уровня вариации в выборках (знаменатель).

F-статистика для однофакторного дисперсионного анализа

$$F = \frac{\text{Межгрупповая вариация}}{\text{Внутригрупповая вариация}}.$$

Число степеней свободы = $k-1$ (числитель) и $n-k$ (знаменатель).

Обратите внимание, что F-статистика имеет *два* значения числа степеней свободы, которые она унаследовала от обоих значений вариации.

В примере с качеством продукции поставщиков F-статистику (с 2 и 55 степеней свободы) вычисляют следующим образом:

$$F = \frac{\text{Межгрупповая вариация}}{\text{Внутригрупповая вариация}} = \frac{269,08}{45,63} = 5,897.$$

Это значение F-статистики показывает, что межгрупповая вариация (обусловленная различиями поставщиков) в 5,897 раза больше внутригрупповой вариации. Таким образом, вариация между поставщиками в 5,897 раза больше, чем можно было бы ожидать, исходя только из вариации отдельных поставщиков. Достаточно ли это много, чтобы утверждать о наличии значимых различий между поставщиками? Для ответа на этот вопрос следует обратиться к статистической таблице.

F-таблица

F-таблица содержит критические значения для распределения F-статистики при условии справедливости нулевой гипотезы, т.е. при истинной нулевой гипотезе значение F-статистики превышает значение из F-таблицы в контролируемом проценте случаев (например, 5%). Чтобы найти критическое значение в таблице, необходимо использовать количество степеней свободы для определения строки F-таблицы и соответствующий уровень проверки (например, 5%) для определения необходимого столбца. Табл. 15.2.2–15.2.5 содержат критические F-значения для уровней проверки 5%, 1%, 0,1% и 10% соответственно.

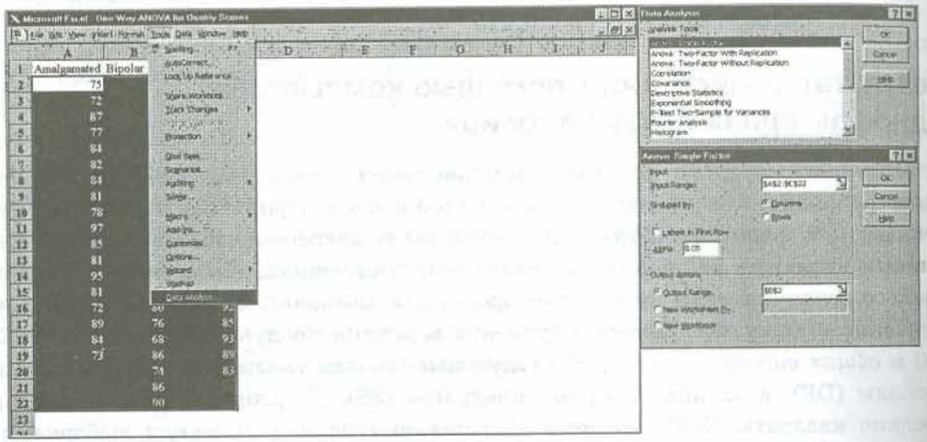
Таблица 15.2.4. Критические F-значения для уровней проверки 0,1%

Число степеней свободы ($n-k$) (знаменатель) для внутригрупповой вариации для однофакторного дисперсионного анализа	Число степеней свободы ($k-1$) (числитель) для межгрупповой вариации для однофакторного дисперсионного анализа																
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
1	405284	500000	540379	562500	576405	585937	592873	598144	602284	605621	610668	615764	620908	626099	631337	633972	636629
2	998,50	999,00	999,17	999,25	999,30	999,33	999,36	999,38	999,39	999,40	999,42	999,43	999,45	999,47	999,48	999,49	999,50
3	167,03	148,50	141,11	137,10	134,58	132,85	131,58	130,62	129,86	129,25	128,32	127,37	126,42	125,45	124,47	123,97	123,47
4	74,137	61,246	56,177	53,436	51,712	50,525	49,658	48,996	48,475	48,053	47,412	46,761	46,100	45,429	44,746	44,400	44,051
5	47,181	37,122	33,202	31,085	29,752	28,834	28,163	27,649	27,244	26,917	26,418	25,911	25,395	24,869	24,333	24,060	23,785
6	35,507	27,000	23,703	21,924	20,803	20,030	19,463	19,030	18,688	18,411	17,989	17,559	17,120	16,672	16,214	15,981	15,745
7	29,245	21,689	18,772	17,198	16,206	15,521	15,019	14,634	14,330	14,083	13,707	13,324	12,932	12,530	12,119	11,909	11,697
8	25,415	18,494	15,829	14,392	13,485	12,858	12,398	12,046	11,767	11,540	11,194	10,841	10,480	10,109	9,727	9,532	9,334
9	22,857	16,387	13,902	12,560	11,714	11,128	10,698	10,368	10,107	9,894	9,570	9,238	8,898	8,548	8,187	8,001	7,813
10	21,040	14,905	12,553	11,283	10,481	9,926	9,517	9,204	8,956	8,754	8,445	8,129	7,804	7,469	7,122	6,944	6,762
12	18,643	12,974	10,804	9,633	8,892	8,379	8,001	7,710	7,480	7,292	7,005	6,709	6,405	6,090	5,762	5,593	5,420
15	16,587	11,339	9,335	8,253	7,567	7,092	6,741	6,471	6,256	6,081	5,812	5,535	5,248	4,950	4,638	4,475	4,307
20	14,818	9,953	8,098	7,095	6,460	6,018	5,692	5,440	5,239	5,075	4,823	4,562	4,290	4,005	3,703	3,544	3,378
30	13,293	8,773	7,054	6,124	5,534	5,122	4,817	4,581	4,393	4,239	4,000	3,753	3,493	3,217	2,920	2,759	2,589
60	11,973	7,767	6,171	5,307	4,757	4,372	4,086	3,865	3,687	3,541	3,315	3,078	2,827	2,555	2,252	2,082	1,890
120	11,378	7,321	5,781	4,947	4,416	4,044	3,767	3,552	3,379	3,237	3,016	2,783	2,534	2,262	1,950	1,767	1,543
∞	10,827	6,908	5,422	4,617	4,103	3,743	3,475	3,266	3,097	2,959	2,742	2,513	2,266	1,990	1,660	1,447	1,000

Таблица 15.2.5. Критические F-значения для уровней проверки 10%

Число степеней свободы ($n-k$) (знаменатель) для внутригрупповой вариации для однофакторного дисперсионного анализа	Число степеней свободы ($k-1$) (числитель) для межгрупповой вариации для однофакторного дисперсионного анализа																
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
1	39,863	49,500	53,593	55,833	57,240	58,204	58,906	59,439	59,858	60,195	60,705	61,220	61,740	62,265	62,794	63,061	63,328
2	8,526	9,000	9,162	9,243	9,293	9,326	9,349	9,367	9,381	9,392	9,408	9,425	9,441	9,458	9,475	9,483	9,491
3	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,216	5,200	5,184	5,168	5,151	5,143	5,134
4	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,896	3,870	3,844	3,817	3,790	3,775	3,761
5	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,268	3,238	3,207	3,174	3,140	3,123	3,105
6	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,905	2,871	2,836	2,800	2,762	2,742	2,722
7	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,668	2,632	2,595	2,555	2,514	2,493	2,471
8	3,458	3,113	2,924	2,806	2,726	2,668	2,624	2,589	2,561	2,538	2,502	2,464	2,425	2,383	2,339	2,316	2,293
9	3,360	3,006	2,813	2,693	2,611	2,551	2,505	2,469	2,440	2,416	2,379	2,340	2,298	2,255	2,208	2,184	2,159
10	3,285	2,924	2,728	2,605	2,522	2,461	2,414	2,377	2,347	2,323	2,284	2,244	2,201	2,155	2,107	2,082	2,055
12	3,177	2,807	2,606	2,480	2,394	2,331	2,283	2,245	2,214	2,188	2,147	2,105	2,060	2,011	1,960	1,932	1,904
15	3,073	2,695	2,490	2,361	2,273	2,208	2,158	2,119	2,086	2,059	2,017	1,972	1,924	1,873	1,817	1,787	1,755
20	2,975	2,589	2,380	2,249	2,158	2,091	2,040	1,999	1,965	1,937	1,892	1,845	1,794	1,738	1,677	1,643	1,607
30	2,881	2,489	2,276	2,142	2,049	1,980	1,927	1,884	1,849	1,819	1,773	1,722	1,667	1,606	1,538	1,499	1,456
60	2,791	2,393	2,177	2,041	1,946	1,875	1,819	1,775	1,738	1,707	1,657	1,603	1,543	1,476	1,395	1,348	1,291
120	2,748	2,347	2,130	1,992	1,896	1,824	1,767	1,722	1,684	1,652	1,601	1,545	1,482	1,409	1,320	1,265	1,193
∞	2,706	2,303	2,084	1,945	1,847	1,774	1,717	1,670	1,632	1,599	1,546	1,487	1,421	1,342	1,240	1,169	1,000

выберите Data Analysis (Анализ данных)[7], а затем операцию ANOVA: Single factor (Однофакторный дисперсионный анализ) и щелкните на кнопке OK. Убедитесь, что в полученном диалоговом окне правильно указан диапазон ячеек с исходными данными, щелкните на кнопке Output Range (Выходной интервал) и определите, в какое место рабочей таблицы вы хотите поместить свои данные. Затем для получения результата щелкните на кнопке OK. Ниже показаны исходная рабочая электронная таблица, диалоговые окна и результаты вычислений для примера о качестве продукции поставщиков. Обратите внимание, что результаты содержат размеры выборки (18, 21 и 19), средние (82,06; 80,67 и 87,68), вариацию между выборками (MS между группами — 269,08), вариацию внутри выборок (MS внутри групп — 45,63), значение F-статистики, 5,897, соответствующее *p*-значение, 0,00478, и критическое F-значение из F-таблицы, равное 3,165.



15.3. Тест наименьшего значимого различия: какие пары различаются?

Что если вы захотите узнать, *какие* выборочные средние значимо отличаются от других? F-тест не даст такой информации, он просто установит наличие или отсутствие различий. Существует ряд различных методов решения этой проблемы. Представленный здесь метод **проверки наименьшего значимого различия** основан на t-тесте для средней разности между парами выборок.

Существует обязательное правило, которое должно выполняться, чтобы вероятность совершения ошибки I рода оставалась на низком уровне, равном 5% (или любом другом выбранном уровне). Проблема состоит в том, что приходится

[7] Если в меню Tools (Сервис) отсутствует пункт Data Analysis (Анализ данных), то сначала убедитесь, что вы выбрали ячейку электронной таблицы (а не график, например). Если вы все же не можете найти Data Analysis (Анализ данных), поищите Add-Ins (Надстройки) и поставьте отметку возле Analysis ToolPak (Пакет анализа). Если это не поможет, то, видимо, необходимо переустановить Excel.

выполнять много t-тестов (по одному для каждой пары выборок) и, несмотря на то, что вероятность *индивидуальной* ошибки для каждого теста остается на уровне 5%, *групповая* ошибка для всех пар может быть значительно выше.[8]

Обязательное требование для тестирования отдельных пар

Если F-тест незначим, не нужно проверять различия отдельных выборок между собой. F-тест уже установил отсутствие значимых различий. Если F-тест незначим, но некоторый t-тест является значимым, F-тест преобладает и, таким образом, t-тест не является действительно значимым.

Если F-тест значим, то можно двигаться дальше и проверять различия для отдельных выборок, чтобы установить, какие именно пары выборок различаются между собой.

При проведении t-теста для принятия решения о различии двух конкретных выборок учитывают следующие показатели.

1. *Разность средних* этих выборок, которая вычисляется путем вычитания одного среднего из другого. (Не имеет значения, что из чего вычитать, пока вы помните, как вы это делали.)
2. *Стандартная ошибка* этой разности средних.
3. *Число степеней свободы*, равное $n - k$, независимо от сравниваемых групп, поскольку стандартная ошибка несет информацию обо всех выборках.
4. Стандартную ошибку вычисляют следующим образом.

Стандартная ошибка разности средних для двух выборок

$$\text{Стандартная ошибка} = \sqrt{(\text{внутривыборочная изменчивость})\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

где n_i и n_j — размеры двух выборок, которые сравниваются.

Обратите внимание, что эта стандартная ошибка может меняться в зависимости от сравниваемых пар, так как изменчивость их выборочных средних зависит от размеров выборок.

Далее, как обычно, с помощью t-таблицы либо определяют доверительный интервал разности средних для генеральных совокупностей, чтобы определить, включает ли этот интервал заданное значение 0 (означающее отсутствие различий), либо вычисляют t-статистику (путем деления средней разности на стандартную ошибку) и сравнивают полученный результат со значением из t-таблицы.

В примере с качеством продукции поставщиков имеются три пары компаний для сравнения: Amalgamated и Bipolar, Amalgamated и Consolidated и Bipolar и Consolidated. Можно ли сравнивать эти пары? Да, потому что F-тест демонстрирует наличие значимых различий средних оценок качества продукции этих поставщиков.

[8] Значение групповой ошибки представляет собой вероятность того, что по крайней мере один из t-тестов неверно установит значимость там, где фактически нет различия между средними генеральных совокупностей.

Ниже приведены вычисления для сравнения компаний Amalgamated и Bipolar с использованием приблизительного значения 1,960 из t-таблицы для $n-k=58-3=55$ степеней свободы.[9]

$$\text{Разность средних} = 80,667-82,056=-1,389.$$

$$\text{Стандартная ошибка} = \sqrt{(\text{внутривыборочная изменчивость})\left(\frac{1}{n_2}+\frac{1}{n_1}\right)} =$$

$$= \sqrt{(45,63)\left(\frac{1}{21}+\frac{1}{18}\right)} = \sqrt{45,63\times 0,103175} = \sqrt{4,708} = 2,170$$

95% доверительный интервал для разности средних генеральных совокупностей находится между

$$-1,389-(1,960\times 2,170)=-5,64 \text{ и } -1,389+(1,960\times 2,170)=2,86.$$

Значение t-статистики равно

$$t=-\frac{1,389}{2,170}=-0,640.$$

Итак, компания-поставщик Bipolar имеет более низкую оценку качества, чем компания Amalgamated, в среднем разность составляет −1,389 пунктов, но эта разность *не является статистически значимой*. Вы на 95% уверены, что значение разности находится в интервале от −5,64 до 2,86. Ввиду того что этот доверительный интервал включает нуль в качестве возможного значения разности, вы принимаете нулевую гипотезу об отсутствии различий в генеральной совокупности между средними оценками качества компаний Amalgamated и Bipolar. Можно также выполнить t-тест, определив, что значение t-статистики (−0,640) меньше по абсолютному значению, чем табличное значение 1,960.

Однако поставщик Consolidated действительно *имеет* значимо более высокое качество продукции, чем поставщики Amalgamated и Bipolar. Ниже приведены вычисления для сравнения компаний Amalgamated с Consolidated.

$$\text{Разность средних} = 87,684-82,056=5,628.$$

$$\text{Стандартная ошибка} = \sqrt{(45,63)\left(\frac{1}{19}+\frac{1}{18}\right)} = 2,222.$$

95% доверительный интервал для разности средних генеральных совокупностей находится между

$$5,628-(1,960\times 2,222)=1,27 \text{ и } 5,628+(1,960\times 2,222)=9,98;$$

$$t=\frac{5,628}{2,222}=2,533.$$

Ниже приведены вычисления для сравнения Bipolar с Consolidated.

[9] t-значение, равное 1,960, соответствует бесконечному числу степеней свободы, и его часто используют для числа степеней свободы 40 и больше. Немного другой ответ будет получен, если взять более точное значение, равное 2,004.

$$\text{Разность средних} = 87,684-80,667=7,017.$$

$$\text{Стандартная ошибка} = \sqrt{(45,63)\left(\frac{1}{19}+\frac{1}{21}\right)} = 2,139.$$

95% доверительный интервал для разности средних генеральных совокупностей находится между

$$7,017-(1,960\times 2,139)=2,82 \text{ и } 7,017+(1,960\times 2,139)=11,21;$$

$$t=\frac{7,017}{2,139}=3,281.$$

Итак, о чем же говорит нам результат дисперсионного анализа качества продукции этих трех поставщиков.

1. Есть значимые различия качества продукции поставщиков. F-тест свидетельствует, что средние значения оценок качества продукции этих трех поставщиков не одинаковы.
2. Consolidated *имеет существенно более высокое качество в сравнении с каждым из двух других поставщиков* (на основе теста наименьшего значимого различия).
3. Два других поставщика, Amalgamated и Bipolar, не имеют значимых различий с точки зрения среднего уровня качества своей продукции.

15.4. Более сложные планы дисперсионного анализа

Если ваши данные представляют собой более сложную структуру, чем просто один набор выборок, дисперсионный анализ может быть преобразован для ответа на более сложные вопросы.

Для того чтобы можно было использовать дисперсионный анализ, данные должны представлять собой набор выборок с одной и той же базовой единицей измерения, как и при обычном однофакторном дисперсионном анализе. Новым является то, что эти выборки могут быть упорядочены в соответствии с некоторой структурой или некоторым образцом. Например, выяснить, существуют ли значимые различия в заработной плате между четырьмя группами "белые мужчины", "белые женщины", "не белые мужчины", "не белые женщины", можно с помощью однофакторного дисперсионного анализа, но двухфакторный дисперсионный анализ позволяет определить различия в заработной плате между людьми разного пола и разной расы.

Все также необходимо обеспечить, чтобы данные удовлетворяли основным допущениям. Во-первых, предполагается, что каждая выборка представляет собой случайную выборку из генеральной совокупности, для которой будет выполняться обобщение. Во-вторых, предполагается, что каждая генеральная совокупность подчиняется нормальному распределению и стандартные отклонения этих генеральных совокупностей равны между собой.

Существует другой способ представления структуры данных, предназначенный для выполнения дисперсионного анализа: необходим многомерный набор данных, в котором в точности одна переменная является количественной (основная измеряемая величина), а все остальные переменные являются качественными. Качественные переменные в совокупности определяют, каким образом количественные наблюдения группируются в выборки.

Разнообразие — вот, что придает вкус жизни

Существует огромное разнообразие в мире ANOVA, обусловленное множеством способов взаимосвязи между выборками. Один вид анализа отличается от другого *планом*, т.е. способом сбора данных. При использовании продвинутых методов ANOVA от вас зависит, чтобы компьютер использовал ANOVA-модель, которая подходит для ваших данных. В большинстве случаев компьютер не может выбрать корректную модель исходя только из данных. Ниже приведены некоторые более продвинутые варианты ANOVA.

Двухфакторный дисперсионный анализ

Если выборки образуют таблицу, в которой одному фактору соответствуют строки, а другому — столбцы, можно поставить три основных вопроса: "Влияет ли первый фактор на появление каких-либо различий?" "Влияет ли второй фактор на появление каких-либо различий?" "Зависит ли влияние первого фактора от второго, или оба фактора действуют независимо друг от друга?" Первые два вопроса имеют отношение к *главным эффектам* каждого из факторов, а третий относится к *взаимодействию* факторов.

Ниже представлен пример, для которого подходит двухфакторный дисперсионный анализ. Первый фактор — *смена*, указывает была ли дневная рабочая смена, ночная смена или пересменка при выпуске партии продукции. Второй фактор — *поставщик*, указывает, какой из четырех поставщиков поставил сырье. Измеряемая величина — *оценка качества* выпущенной продукции. Первый вопрос касается главного эффекта фактора рабочей смены: значимо ли различаются оценки качества продукции, выпущенной в разных сменах? Второй вопрос касается главного эффекта фактора поставщика: значимо ли различается качество выпускаемой продукции, выпущенной из сырья разных поставщиков? Третий вопрос относится к взаимодействию факторов смены и поставщика, например: можно ли утверждать, что оценки качества продукции в трех сменах изменяются по-разному, в зависимости от того, сырье какого из поставщиков использовалось?

В качестве конкретного примера взаимодействия факторов предположим, что качество продукции, изготовленной в ночную смену, обычно выше, чем изготовленное в другие смены, но ночная бригада испытывает беспокойство при работе с сырьем одного из поставщиков (другие смены такого беспокойства не проявляют). Здесь взаимодействие заключается в том, что сырье поставщика влияет на работу смен по-разному. Если бы такого взаимодействия не было, то *все* смены испытывали бы одинаковые проблемы при работе с сырьем данного поставщика.

На рис. 15.4.1 показано, как можно представить взаимодействие факторов. Если на одной оси отразить виды смен (дневная, ночная, пересменка), а на другой — среднее значение оценок качества продукции, то можно провести линию для каждого из поставщиков (A, B, C). Разный угол наклона линий указывает на взаимодействие факторов.

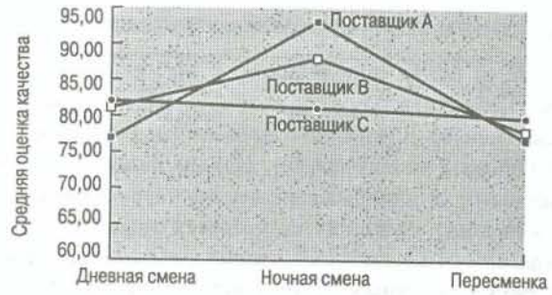


Рис. 15.4.1. График средних демонстрирует взаимодействие факторов — линии имеют разный угол наклона как при движении вниз, так и при движении вверх. Обратите внимание, в частности, что ночная смена (в середине) обычно выпускает продукцию лучшего качества, чем другие смены, за исключением случая, когда она работает с сырьем поставщика C

На рис. 15.4.2 показано, как выглядел бы тот же график, если бы факторы абсолютно не взаимодействовали.



Рис. 15.4.2. Вид графика в случае абсолютного отсутствия взаимодействия. Обратите внимание, что угол наклона линий одинаков как при движении вниз, так и при движении вверх. В этом случае качество продукции, выпускаемой в ночную смену, выше, независимо от поставщика сырья (будьте внимательны, сравнивая качество продукции, выпускаемой из сырья одного поставщика). Заметим также, что все смены имеют одинаковые проблемы с сырьем поставщика C

Конечно, в реальной жизни в данных всегда присутствует случайность, так что почти всегда есть некоторое взаимодействие. Цель теста значимости взаимодействия факторов в дисперсионном анализе заключается в том, чтобы проверить, является ли наблюдаемое взаимодействие значимым (т.е. отличается в статистическом смысле более чем просто случайно от ситуации "отсутствия взаимодействия").

Три фактора и более

При наличии трех и более определяющих выборку факторов дисперсионный анализ по-прежнему позволяет изучить *главный эффект* каждого фактора (чтобы установить, оказывает ли он влияние) и *взаимодействие* факторов (чтобы установить, как факторы связаны между собой). Новым является то, что мы имеем дело с большим, по сравнению с двухфакторным дисперсионным анализом, количеством видов взаимодействий, каждый из которых изучается отдельно. Здесь рассматриваются *двухфакторные взаимодействия* различных пар факторов, трехфакторные взаимодействия различных троек факторов и т.д., пока не будет рассмотрено взаимодействие высшего уровня, включающее все факторы сразу.

Ковариационный анализ, ANCOVA

Ковариационный анализ объединяет регрессионный и дисперсионный анализы. Например, в дополнение к основным данным ANOVA у вас может быть важная дополнительная количественная переменная. Вместо того чтобы выполнять дисперсионный анализ, игнорируя эту дополнительную переменную, или выполнять регрессионный анализ, игнорируя группы, можно применить ковариационный анализ, который учитывает и то и другое. Можно рассматривать ковариационный анализ либо с точки зрения отношений между отдельными регрессиями, построенными для каждой из выборок, либо с точки зрения дисперсионного анализа, который скорректирован с учетом различий, внесенных дополнительной переменной.

Многомерный дисперсионный анализ (MANOVA)

При наличии более одной количественной зависимой переменной можно использовать многомерный дисперсионный анализ, чтобы изучить различия всех этих зависимых переменных между выборками. Если, например, имеется три количественных показателя готовой продукции (внешний вид, качество работы, уровень шума при работе), то можно применить MANOVA, чтобы посмотреть, значимо ли различаются эти величины под влиянием главного эффекта фактора смены (дневной, ночной, пересменки) и фактора поставщика.

Как читать таблицу ANOVA

Общий вид традиционной ANOVA таблицы приведен в табл. 15.4.1.

Несмотря на то что такую таблицу полезно использовать при проверке гипотезы о средних генеральных совокупностей (если вам известно, как ею пользоваться), она имеет два серьезных недостатка. Во-первых, в таблице отсутствуют исходные данные. Это означает, что необходима отдельная таблица, в которой представлены средние значения, например качество продукции в ночную и дневную смены. Во-вторых, большая часть такой таблицы не имеет прямой практической интерпретации: для многих применений полезны только первый (источник вариации) и последний (p-значение) столбцы. Остальные столбцы представляют собой лишь промежуточные этапы вычислений для получения p-значений, которые и являются результатом проверки статистической значимости. Тем не менее для обоснования утверждений о статистической значимости в дисперсионном анализе традиционно используют именно эту таблицу.

Таблица 15.4.1. Общий вид традиционной таблицы ANOVA

Источник вариации	Сумма квадратов	Степени свободы	Средний квадрат	F-значение	p-значение
Источник 1	SS_1	df_1	$MS_1 = SS_1 / df_1$	$F_1 = MS_1 / MS_e$	p_1
Источник 2	SS_2	df_2	$MS_2 = SS_2 / df_2$	$F_2 = MS_2 / MS_e$	p_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
Источник k	SS_k	df_k	$MS_k = SS_k / df_k$	$F_k = MS_k / MS_e$	p_k
Ошибка или остаток	SS_e	df_e	$MS_e = SS_e / df_e$		

Каждую гипотезу проверяют с помощью F-теста, в ходе которого сравнивают значение среднего квадрата для данного источника вариации (которое является большим, если этот источник вариации приводит к различиям в значениях количественных показателей) со средним квадратом ошибки, выясняя при этом, насколько сильнее влияние данного источника вариации в сравнении с чистой случайностью. Чтобы определить, значимо ли влияние этого источника вариации (скажем, i-го), сравнивают вычисленное F-значение (F_i) с табличным F-значением, взятым с числом степеней свободы для этого источника (df_i) в числителе и числом степеней свободы для ошибки (df_e) в знаменателе. Или же следует просто посмотреть на p-значение (p_i) и принять решение о том, что результат "значим", если это значение достаточно мало, например $p < 0,05$.

Пример. *Влияние изменений цен и вида продукции на объемы продаж бакалейных товаров*

Мы ожидаем увеличения объема продаж в то время, когда продукция продается по цене ниже обычной. Если продукция может храниться дома у потребителя, то для такой продукции следует ожидать большего увеличения объема продажи, чем для более скоропортящейся или менее часто потребляемой. Такие вопросы рассматривались в исследовании, проведенном Литвак, Калантоне и Уаршоу (Litvack, Calantone and Warshaw).[10] Они использовали двухфакторный дисперсионный анализ со следующей базовой структурой.

Первый фактор определяет два типа товаров: те, которые могут храниться, и те, которые не могут храниться. К первым относятся такие, которые покупатель может купить впрок и хранить у себя дома, например еда для собаки, салфетки, рыбные консервы. Ко вторым можно отнести горчицу, сыр, овсяные завтраки.

Второй фактор определяется тремя уровнями цены: сниженная на 20%, без изменения и повышенная на 20% по сравнению с обычной ценой в этом магазине.

Измеряется изменение объема продаж, выраженное количеством проданных единиц товара на один миллион долларов объема продаж бакалейных товаров в каждом магазине. Обратите внимание, что такое определение изменения продаж учитывает разницу в объемах продаж бакалейных товаров для различных магазинов и позволяет анализировать совместно большие и малые бакалейные магазины.

Изменение продаж измерялось для разных продуктов каждого типа и при разных уровнях цен. Полученные при этом результаты приведены в ANOVA табл. 15.4.2.

[10] Litvack D. S., Calantone R. J. and Warshaw P. R. "An Examination of Short-Term Retail Grocery Price Effects", *Journal of Retailing*, 61 (1985), p. 9–25.

Таблица 15.4.2. ANOVA-таблица для установления влияния типа товара и цены на объемы продаж

Источник вариации	Сумма квадратов	Степени свободы	Средний квадрат	*F*-значение	*p* -значение
Тип товара	0,469	1	0,469	0,32	0,5694
Изменение цены	33,284	2	16,642	11,50	0,0001
Взаимодействие:					
Тип товара × изменение цены	13,711	2	6,856	4,74	0,0095
Ошибка или остаток	77,579	261	1,447		

Указанное для типа товара *p* -значение 0,5694 свидетельствует об *отсутствии значимых различий* между средними значениями объемов продаж для тех товаров, которые могут храниться, и тех, которые не могут. Это отчасти удивительно, так как мы ожидали, что разница будет. Однако посмотрим далее. *p* -значение, 0,0001, для фактора изменения цены показывает, что имеет место *очень высоко значимая разница* между средними значениями объемов продаж по пониженной, повышенной и обычной цене. Таким образом, изменение цены оказывает значимое влияние на объем продаж.

Результат взаимодействия высоко значим ($p = 0{,}0095$, что меньше 0,01). Это свидетельствует о том, что изменение объема продаж в ответ на изменение цены зависит от типа товара (может товар храниться или не может). Иными словами, сектор торговли товарами, которые могут храниться, иначе реагирует на изменение цен, чем сектор торговли другими товарами.

Почему же возможно, что главный эффект влияния на объем продаж такого фактора, как тип товара, незначим, в то время как эффект взаимодействия типа товара и изменения цены оказался значим? Вспомним, что главный эффект учитывает только среднее значение объема продаж для каждого типа товаров, в то время как эффект взаимодействия охватывает все комбинации типа товара и изменения цены.

Хотя таблица ANOVA содержит все необходимые данные для определения значимости результата проверки, много важной информации отсутствует. В частности, какое влияние на объем продажи пригодного для хранения товара оказывает 20% понижение цены? Ответ на такой практически важный вопрос в таблице ANOVA мы не найдем. Чтобы ответить на него, необходимо изучить средние значения объемов продаж, приведенные в табл. 15.4.3 и показанные на рис. 15.4.3.

Каждое среднее представляет 12 наименований товаров и 4 магазина. Обратите внимание, что объемы продаж товаров, которые могут храниться (слева на рисунке), *не* одинаково располагаются выше или ниже объемов продаж товаров, которые не могут хранится (расположены на рисунке справа). Это помогает объяснить, почему незначим главный эффект такого фактора, как тип товара. Заметим также, что объемы продаж товаров, которые могут храниться, возрастают только при понижении цены. Это происходит только при таком сочетании этих двух факторов, что является частью эффекта взаимодействия, который, в свою очередь, является значимым. Также ясно, что объемы продаж падают при повышении цены и достигают наибольшего значения при самых низких ценах, следствием чего является значимость главного эффекта для фактора изменения цены.

Таблица 15.4.3. Процентное изменение стандартизированных объемов продаж

	Изменение цены		
Тип товара	цена, сниженная на 20%	цена без изменения	цена, повышенная на 20%
Может храниться	54,95	1,75	–24,10
Не может храниться	10,55	6,95	–7,60

Итак, не думайте, что таблица ANOVA (такая как табл. 15.4.2) раскрывает вам всю картину. Если у вас остаются вопросы, смотрите на значения средних (как показано в табл. 15.4.3 и на рис. 15.4.3), чтобы понять, что происходит в действительности.



Рис. 15.4.3. Среднее изменение объемов продаж для шести комбинаций типа товара и изменения цены. В ANOVA-таблице приведены результаты проверки гипотез о шести средних значениях объемов продаж для генеральных совокупностей, которые эти средние представляют

Пример. *Шутки на рабочем месте*

Какой вид шуток недопустим на рабочем месте? Почему одни люди обижаются на некоторые виды шуток, а другие нет? Эти проблемы с помощью дисперсионного анализа изучали Шмельтцер и Лип (Smeltzer и Leap).[11] Как руководителя вас эти проблемы могут интересовать независимо от вашего личного отношения к юмору, потому что "шутки могут иметь своим результатом судебные разбирательства относительно гражданских и человеческих прав, а также оказывать влияние на качество работы".

Исследование было проведено с помощью трехфакторного дисперсионного анализа. В качестве трех факторов рассматривались пол (мужской или женский), раса (белая или черная) и опыт работы (неопытный, т.е. со стажем работы менее одного года, или опытный). Рассмотрение всех комбинаций этих трех факторов, каждый из которых имеет две категории, позволяет выделить восемь различных типов служащих (черный мужчина без опыта работы, черный мужчина с опытом работы, белый мужчина без опыта работы и т.д.). Давайте посмотрим, как 165 служащих оценили 5 шуток на сексуальную тему по 7-балльной шкале неуместности подобной шутки на рабочем месте.

Результаты ANOVA приведены в табл. 15.4.4.

Нет необходимости приводить суммы квадратов, средние значения квадратов и значения остатков. Проверки всех обычно используемых гипотез можно выполнить на основе приведенных *p* -значений.

Что касается главных эффектов, то только эффект фактора расы оказался значимым. Это говорит о том, что люди разной расы в целом по-разному относятся к уместности шуток на сексуальные темы на рабочем месте. Как обычно, ANOVA-таблица не показывает, какие именно группы считают такие шутки более уместными на рабочем месте. Чтобы ответить на этот вопрос, необходимо изучить средние значения оценок, выставленных служащими каждой из групп. Различия невелики (5,4 балла — для белых и 4,36

[11] Smeltzer L. R. and Leap T. L. "An Analysis of Individual Reactions to Potentially Offensive Jokes in Work Settings", *Human Relations*, 41 (1988), p. 295–304.

балла — для черных в этом исследовании), но вероятность того, что эта разница обусловлена только фактором случайности, очень мала.

Среди двухфакторных взаимодействий значимым является только взаимодействие пол × раса. Это свидетельствует о том, что различия в отношении к такого рода шуткам между мужчинами и женщинами зависит от их расы. Трехфакторное взаимодействие незначимо, что свидетельствует об отсутствии более детальных различий в отношении к подобным шуткам, на которые можно было бы указать исходя из имеющегося набора данных.

Таблица 15.4.4. ANOVA-таблица для определения уместности шуток на сексуальную тему на рабочем месте

Источник вариации	Степени свободы	*F*-значение	*p* -значение
Главные эффекты			
Пол	1	2,83	0,09
Раса	1	15,59	0,0001
Опыт	1	0,54	0,46
Двухфакторные взаимодействия			
Пол × раса	1	6,87	0,009
Пол × опыт	1	0,00	1,0
Раса × опыт	1	2,54	0,11
Трехфакторное взаимодействие			
Пол × раса × опыт	1	1,44	0,23

15.5. Дополнительный материал

Резюме

Дисперсионный анализ (сокращенно **ANOVA**) задает общую схему проверки статистических гипотез, основанную на тщательном изучении различных источников вариации в сложной ситуации. Для проверки каждой из гипотез в дисперсионном анализе используют **F-тест**, основанный на **F-статистике**, которая представляет собой отношение двух дисперсий. Числитель в таком отношении представляет собой вариацию, обусловленную конкретным интересующим нас эффектом, который мы и проверяем, а знаменатель — вариацию, обусловленную случайностью. Если это отношение больше табличного F-значения, эффект значим. **Однофакторный дисперсионный анализ**, в частности, используют для проверки значимости различий между собой значений средних, характеризующих различные ситуации.

Не забывайте о первичном разведывательном анализе данных. Блочная диаграмма поможет сравнить сразу несколько распределений, что позволит вам увидеть структуру данных, определить проблемы (если они есть) и проверить начальные условия, необходимые для дисперсионного анализа, а именно нормальность распределений и равенство вариации.

Данные для однофакторного дисперсионного анализа представляют собой k независимых одномерных выборок, элементы которых измерены в одинаковых единицах. Однофакторный дисперсионный анализ сравнивает два следующих источника вариации:

- межгрупповая вариация (между выборками);
- внутригрупповая вариация (внутри каждой выборки).

Для корректного применения однофакторного дисперсионного анализа необходимо выполнение двух условий.

1. Набор данных состоит из k случайных выборок из k генеральных совокупностей.
2. Каждая генеральная совокупность подчиняется нормальному распределению, и стандартные отклонения во всех генеральных совокупностях одинаковы, т.е. $\sigma_1 = \sigma_2 = \ldots = \sigma_k$.
3. Нулевая гипотеза утверждает, что между генеральными совокупностями нет различий, а альтернативная гипотеза утверждает, что некоторые различия имеют место:
   - $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (все средние равны между собой);
   - $H_1 : \mu_i \neq \mu_j$ по крайней мере для одной пары генеральных совокупностей (*не* все средние равны между собой).

Общее (главное) среднее представляет собой среднее всех значений из всех выборок:

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n} = \frac{\sum_{i=1}^{k} n_i\bar{X}_i}{n}.$$

Межгрупповая (межвыборочная) вариация измеряет различие выборочных средних, а **внутригрупповая (внутривыборочная) вариация** измеряет вариабельность каждой из выборок:

$$\text{межгрупповая вариация} = \frac{n_1\left(\bar{X}_1 - \bar{X}\right)^2 + n_2\left(\bar{X}_2 - \bar{X}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{X}\right)^2}{k-1} = \frac{\sum_{i=1}^{k} n_i\left(\bar{X}_i - \bar{X}\right)^2}{k-1}.$$

Число степеней свободы $= k-1$.

$$\text{Внутригрупповая вариация} = \frac{(n_1-1)(S_1)^2 + (n_2-1)(S_2)^2 + \ldots + (n_k-1)(S_k)^2}{n-k} = \frac{\sum_{i=1}^{k}(n_i-1)(S_i)^2}{n-k}.$$

Число степеней свободы $= n-k$.

F-статистика представляет собой отношение этих двух значений вариации и показывает меру различия выборочных средних (числитель) по отношению к общему уровню вариации выборок (знаменатель).

$$F = \frac{\text{Межгрупповая вариация}}{\text{Внутригрупповая вариация}}.$$

Число степеней свободы = $k - 1$ (для числителя)
и $n - k$ (для знаменателя).

F-таблица содержит критические значения для распределения F-статистики. Таким образом, если справедлива нулевая гипотеза, то значение F-статистики превышает значение из F-таблицы в контролируемом проценте случаев (например, 5%). F-тест выполняют, сравнивая значение F-статистики (рассчитанное из данных) с критическим значением из F-таблицы.

F-тест определяет только наличие или отсутствие различий. Если F-тест фиксирует значимость различий, то **тест наименьшего значимого различия** используется для сравнения каждой пары выборок, чтобы определить, значимо ли различаются они между собой. Этот тест основан на средней разности между двумя сравниваемыми группами, стандартной ошибке этой разности и количестве степеней свободы ($n - k$):

$$\text{Стандартная ошибка} = \sqrt{(\text{внутригрупповая вариация})\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

где n_i и n_j — размеры двух сравниваемых выборок.

Существует много более сложных методов дисперсионного анализа, включая двухфакторный план и планы более высокого порядка. Для того чтобы любой из таких методов дисперсионного анализа можно было применить, данные должны состоять из набора выборок, представляющих результаты измерения в одних и тех же единицах, как и при обычном однофакторном дисперсионном анализе. Помните, что ANOVA-таблица не раскрывает полностью всю картину; всегда необходимо смотреть на значения средних, чтобы понять, что действительно происходит.

Основные термины

- Дисперсионный анализ (analysis of variance) (ANOVA), 808
- F-тест (F test), 809
- F-статистика (F statistic), 809
- Однофакторный дисперсионный анализ (one-way analysis of variance), 809
- Общее (главное) среднее (grand average), 815
- Межгрупповая вариация (between-sample variability), 815
- Внутригрупповая вариация (within-sample variability), 816
- F-таблица (F table), 817
- Тест наименьшего значимого различия (least-significant-difference test), 824

Контрольные вопросы

1. Объясните, в каком смысле дисперсионный анализ позволяет исследовать дисперсию и, в частности, какие дисперсии анализируются и почему?
2. а) Данные какого типа следует анализировать с помощью однофакторного дисперсионного анализа?

   б) Почему не следует использовать t-тест для независимых выборок вместо однофакторного дисперсионного анализа?
3. Назовите и прокомментируйте два источника вариации в однофакторном дисперсионном анализе.
4. Какое условие должно выполняться, чтобы данные были репрезентативными для генеральной совокупности?
5. Должны ли быть равны размеры выборок в однофакторном дисперсионном анализе?
6. а) Сформулируйте гипотезы однофакторного дисперсионного анализа.

   б) Является ли альтернативная гипотеза очень конкретной в отношении характера различий?
7. Опишите и приведите формулу для каждой из указанных ниже величин, которые используются в однофакторном дисперсионном анализе.

   а) Общий размер выборки, n.

   б) Общее (главное) среднее, $\overline{X}$.

   в) Межгрупповая вариация и соответствующее число степеней свободы.

   г) Внутригрупповая вариация и соответствующее число степеней свободы.

   д) F-статистика и соответствующее число степеней свободы.

   е) F-таблица.
8. Когда можно использовать тест наименьшего значимого различия для сравнения отдельных пар выборок? Когда этого делать нельзя?
9. Почему стандартная ошибка средней разности может принимать различные значения, в зависимости от того, какие выборки сравнивают?

Задачи

1. Была осуществлена проверка трех рекламных акций. В каждом случае использовались разные случайные выборки потребителей из одного и того же города. Оценки характеризуют эффективность рекламы; результаты приведены в табл. 15.5.1.

Таблица 15.5.1. Анализ эффективности рекламы

	Реклама 1	Реклама 2	Реклама 3
Среднее	63,2	68,1	53,5
Стандартное отклонение	7,9	11,3	9,2
Размер выборки (потребителей)	101	97	105

а) Какая реклама оказалась наиболее, а какая наименее эффективной?

б) Определите общий объем выборки, n, общее среднее, $\overline{X}$, и число выборок k.

в) Вычислите межгрупповую вариацию и число степеней свободы для нее.

г) Вычислите внутригрупповую вариацию и число степеней свободы для нее.

2. Выполните следующие задания с использованием данных задачи 1.

а) Определите F-статистику и число степеней свободы для нее.

б) Прокомментируйте F-статистику с точки зрения того, во сколько раз более изменчив один источник вариации в сравнении с другим.

в) Найдите в F-таблице критическое значение для уровня 5%.

г) Каковы результаты F-теста при уровне значимости 5%?

д) Обобщите полученные результаты и сделайте вывод о различиях в эффективности этих рекламных акций для потребителей в этом городе в целом.

3. Выполните следующие задания с использованием данных из задачи 1.

а) Найдите в F-таблице критическое значение для уровня 0,1% и опишите результаты F-теста на этом уровне.

б) Обобщите полученные результаты и сделайте вывод о различиях в эффективности этих рекламных акций для потребителей в этом городе в целом.

4. Выполните следующие задания с использованием данных из задачи 1.

а) Определите среднюю разность между эффективностью рекламы 1 и эффективностью рекламы 2 (вычисляется как эффективность рекламы 2 минус эффективность рекламы 1).

б) Определите стандартную ошибку для этой средней разности.

в) Определите число степеней свободы для этой стандартной ошибки?

г) Определите 99,9% доверительный интервал для средней разности эффективности рекламы 1 и рекламы 2 в генеральной совокупности?

д) Является ли различие оценок эффективности рекламы 1 и рекламы 2 очень высоко значимым? Как вы определили?

5. Выполните приведенные ниже задания на основании данных из задачи 1.

а) Определите среднюю разность и ее стандартную ошибку для *каждой* пары реклам (которые вычисляют как реклама 2 – реклама 1, реклама 1 – реклама 3 и реклама 2 – реклама 3).

б) Выполните проверку для каждой пары реклам на уровне 1% и представьте результаты.

6. Три компании пытаются продать вам свои добавки, уменьшающие отходы химического производственного процесса. Вы не уверены, что их продукты вам подойдут, так как ваш технологический процесс отличается от стандартного в данной отрасли (этот процесс является вашим ноу-хау). Вы бесплатно получили от каждой компании небольшое количество добавок, чтобы испытать их в производстве. В табл. 5.5.2 показаны объемы отходов при использовании каждой из добавок и при прочих одинаковых условиях.

Таблица 15.5.2. Анализ количества отходов

	Компании		
	Sludge Away	Cleen Up	No Yuk
Среднее	245,97	210,92	240,45
Стандартное отклонение	41,05	43,52	35,91
Объем выборки (партии продукции)	10	10	10

а) При использовании какой добавки количество отходов оказалось наибольшим? А при использовании какой добавки наименьшим?

б) Определите общий объем выборки, n, общее среднее, $\overline{X}$, и число выборок k.

в) Вычислите межгрупповую вариацию и ее степени свободы.

г) Вычислите внутригрупповую вариацию и ее степени свободы.

7. Выполните следующие задания на основании данных из задачи 6.

а) Вычислите F-статистику и количество ее степеней свободы.

б) Прокомментируйте F-статистику с точки зрения того, во сколько раз более изменчив один источник вариации в сравнении с другим.

в) Найдите в F-таблице критические значения для уровня 5%.

г) Опишите результаты F-теста на уровне 5%.

д) Обобщите полученные результаты с точки зрения сравнения способности этих трех добавок снижать количество отходов.

8. Допустимо ли в задаче 6 использовать тест наименьшего значимого различия для определения того, уменьшает ли добавка компании Cleen Up количество отходов значимо больше, чем добавка компании Sludge Away (на уровне 5%)? Почему?

9. Выполните следующие задания, используя данные из задачи 6.

а) Найдите в F-таблице критическое значение для уровня 10% и опишите результаты F-теста на этом уровне.

б) Обобщите полученные результаты F-теста с точки зрения сравнения способности этих трех добавок снижать количество отходов.

10. Из данных задачи 6 отберите две добавки с наибольшей средней разницей способности снижения количества отходов и ответьте на следующие вопросы. (Для этой задачи используйте тест наименьшего значимого различия, вычитая меньшее значение из большего даже тогда, когда вы чувствуете, что этого делать не следует.)

а) Определите значение средней разности для этой пары.

б) Определите стандартную ошибку этой средней разности.

в) Сколько у этой стандартной ошибки степеней свободы?

г) Определите двусторонний 90% доверительный интервал для средней разности.

д) Исходя из средней разности, стандартной ошибки, степеней свободы и t-таблицы решите, значимо ли на уровне 10% различаются между собой эти две добавки?

е) Можете ли вы утверждать, что две добавки действительно значимо отличаются на уровне 10%? Почему? (Будьте внимательны, решая эту задачу. Вам могут понадобиться результаты F-теста из предыдущей задачи.)

11. Чтобы лучше распределять свое рабочее время, вы провели небольшое исследование, фиксируя время, затраченное на каждый телефонный звонок (в минутах) в течение одного рабочего дня. Перед тем как внести изменения в организацию своей работы (например, переадресовать некоторые звонки своим подчиненным), вы хотите разобраться в этой ситуации. Продолжительность звонков, сгруппированных по определенным темам, приведена в табл. 15.5.3.

а) Постройте в одном масштабе блочные диаграммы для этих четырех видов звонков и опишите полученную структуру.

б) Вычислите среднее и стандартное отклонения для каждого вида телефонных звонков.

в) Какой из типов звонков имеет наибольшую среднюю продолжительность? А какой наименьшую?

Таблица 15.5.3. Продолжительность телефонных звонков

Информация	Продажи	Обслуживание	Остальные
0,6	5,1	5,2	6,3
1,1	1,7	2,9	1,2
1,0	4,4	2,6	3,1
1,9	26,6	1,2	2,5
3,8	7,4	7,0	3,0
1,6	1,4	14,2	2,6
0,4	7,0	8,4	0,8
0,6	3,9	0,6	
2,2	3,1	26,7	
12,3	1,2	7,7	
4,2	1,9	4,8	
2,8	17,3	7,2	
1,4	7,8	2,7	
	4,3	3,4	
	3,4	13,3	
	1,3		
	2,0		

г) Выполняется ли для данного набора данных необходимое для однофакторного дисперсионного анализа предположение о нормальном распределении и равной вариации? Почему?

д) Вычислите для каждого значения натуральный логарифм и постройте блочные диаграммы для этих логарифмов.

е) Можно ли сказать, что предположение о равной вариации лучше выполняется для логарифмов, чем для исходных данных?

12. Выполните следующие задания, используя данные из задачи 11 (используйте значения логарифмов продолжительности телефонных звонков).

а) Определите общий объем выборки, n, общее среднее, $\overline{X}$, и число выборок k.

б) Определите межгрупповую вариацию и число степеней свободы для нее.

в) Определите внутригрупповую вариацию и число степеней свободы для нее.

13. Выполните следующие задания на основании данных задачи 11 (используйте значения логарифмов продолжительности телефонных звонков).

а) Определите F-статистику и число степеней свободы для нее.

б) Найдите в F-таблице критическое значение для уровня 5%.

в) Опишите результаты F-теста для уровня 5%.

г) Обобщите полученные результаты о различиях между этими четырьмя видами телефонных звонков.

14. Выполните следующие задания с использованием данных из задачи 11 (используйте значения логарифмов продолжительности телефонных звонков).

а) Определите среднюю разность и ее стандартную ошибку для каждой пары типов звонков (в каждом случае вычитая из большего значения меньшее).

б) Какая пара типов телефонных звонков существенно отличается друг от друга (в терминах среднего значения логарифма продолжительности звонка)?

15. Для проверки, значимо ли различается качество продукции различных поставщиков (данные приведены в табл. 15.5.1), используйте вместо однофакторного дисперсионного анализа (ANOVA) множественную регрессию с индикаторными переменными. (Вы можете снова обратиться к материалу об индикаторных переменных в главе 12.)

а) Создайте переменную Y, перечислив все оценки качества в одном длинном столбце. Проделайте это, располагая сначала оценки продукции Amalgamated, затем Bipolar и Consolidated.

б) Создайте две индикаторные переменные, одну — для компании Amalgamated и другую — для компании Bipolar.

в) Выполните множественный регрессионный анализ.

г) Сравните значение F-статистики, полученное из регрессионного анализа со значением F-статистики, полученным из однофакторного дисперсионного анализа. Прокомментируйте полученный результат.

Таблица 15.5.4. Средние оценки качества продукции и таблица ANOVA

	Дневная смена	Ночная смена	Пересменка	Среднее
Поставщик A	77,06	93,12	77,06	82,42
Поставщик B	81,14	88,13	78,11	82,46
Поставщик C	82,02	81,18	79,91	81,04
Среднее	80,08	87,48	78,36	81,97

Дисперсионный анализ качества продукции

Источник вариации:	DF	SS	MS	F	P
смена	2	704,07	352,04	11,93	0,000
поставщик	2	19,60	9,80	0,33	0,720
смена × поставщик	4	430,75	107,69	3,65	0,014
Ошибка	36	1062,05	29,50		
Итого	44	2216,47			

д) Сравните коэффициент регрессии для индикаторных переменных со средними значениями разностей оценок качества для разных поставщиков. Прокомментируйте полученный результат.

е) Дают ли эти два метода — множественная регрессия и однофакторный дисперсионный анализ — разные результаты, или результаты полностью совпадают? Как вы думаете, почему это именно так?

16. В табл. 15.5.4 приведены средние значения оценок качества продукции, усредненные в отношении поставщика исходных материалов (A, B C), и смены (дневная, ночная, пересменка), изготовившей эту партию. Структура представленных данных соответствует компьютерной версии таблицы ANOVA. При проведении этого эксперимента взяли по 5 наблюдений для каждой комбинации значений (комбинации поставщика и смены). Обратите внимание, что в последнем ряду (и столбце) представлены средние значения величин ряда и столбцы соответственно (например, 82,42 — это среднее значение оценки качества продукции, полученное из сырья поставщика A во всех сменах).

а) Сравните среднее для поставщика A со средним для поставщиков B и C. Большая ли (более 2 или 3 единиц) между ними разница?

б) Значимо ли различаются средние значения оценок качества для разных поставщиков? Как вы это определили?

17. Сравните среднее значение качества продукции, изготовленной дневной сменой со средним значением качества продукции, изготовленной в ночную и промежуточную смены (исходя из данных табл. 15.5.4). Большая ли (более 2 или 3 единиц) между ними разница? Значима ли эта разница? Как вы это определили?

Таблица 15.5.5. Влияние на взаимодействие атмосферы конкуренции/сотрудничества и различий во взглядах на ценности

Источник вариации	Сумма квадратов	Степени свободы	Средний квадрат	*F*-значение	*p*-значение
Конкуренция – сотрудничество (A)	3185,77	1	3185,77	4,00	0,049682
Разногласия (B)	58,04	1	54,04	0,07	0,792174
A × B	424,95	1	424,95	0,53	0,469221
Ошибка	51 729,98	65	795,85		
Итого	55 370,49	68			

18. Существует ли значимое взаимодействие поставщика и смены (исходя из данных табл. 15.5.4)? Обоснуйте и поясните ваш ответ.

19. Что лучше: конкуренция или сотрудничество? И зависит ли ответ от того, разделяют ли участники одинаковые ценности? Исследование Козиера и Дальтона (Cosier and Dalton) проливает свет на этот вопрос.[12] Одна из их таблиц ANOVA составила основу для табл. 15.5.5.

а) Среднее значение результата выше у сотрудничающей группы, чем у конкурирующей. Значимо ли оно выше? Как вы это определили?

б) Среднее значение результата выше, если разногласий меньше. Оказывают ли расхождения в ценностях значимое влияние на результат? Почему вы так считаете?

в) Является ли взаимодействие факторов существенным? О чем это говорит?

20. Действительно ли цены в универмагах выше цен в магазинах производителя? Кирби и Дардис (Kirby and Dardis)[13] изучали на протяжении 13 недель цены 20 наименований товаров (рубашки, брюки и т.д.) и обнаружили, что цены в универмагах действительно на 40% выше цен в магазинах производителя. В табл. 15.5.6 представлена несколько сокращенная ANOVA-таблица, взятая из их отчета.

а) Действительно ли высокие цены (в среднем на 40% выше в универмагах) значимо выше? Почему?

б) Какой вид дисперсионного анализа здесь использован?

в) Определите три фактора в этом анализе. Сколько категорий имеет каждый из факторов?

г) О чем свидетельствует *p*-значение для главного эффекта фактора B?

д) Значимо ли различаются цены разных недель? Как это узнать?

[12] Cosier R. A. and Dalton D. R. "Competition and Cooperation: Effects of Value Dissensus and Predisposition to Help", *Human Relations*, 41 (1988), p. 823–839.

[13] Kirby G. H. and Dardis R. "Research Note: A Pricing Study of Women's Apparel in Off-Price and Department Stores", *Journal of Retailing*, 62 (1986), p. 321–330.

Таблица 15.5.6. Дисперсионный анализ влияния типа магазина, вида товара и недели на цену

Источник вариации	Сумма квадратов	Степени свободы	Средний квадрат	F-значение	p-значение
Тип магазина (A)	1794577789	1	1794577789	1121,52	0,000000
Вид товара (B)	25726794801	19	1354041	864,21	0,000000
Неделя (C)	246397563	12	2053313	12,83	0,000000
Двухфакторные взаимодействия:					
A × B	970172936	19	510617	31,91	0,000000
A × C	69197628	12	5766469	3,60	0,000027
B × C	320970292	228	140776	0,88	0,884253
Трехфакторное взаимодействия					
A × B × C	264279428	228	115912	0,72	0,998823
Остаток	1664128185	1040	1600123		
Итого	31056518626	1559			

е) Рассмотрите взаимодействие типа магазина и вида товара. Значимо ли оно? О чем это говорит?

ж) Рассмотрите взаимодействие типа магазина и недели. Значимо ли оно? О чем это говорит?

з) Рассмотрите взаимодействие вида товара и недели. Значимо ли оно? О чем это говорит?

и) Обычно мы изучаем p-значения, чтобы посмотреть, достаточно ли они малы, чтобы можно было заявить, что результат является статистически значимым. Однако для трехфакторного взаимодействия p-значение подозрительно велико, что наводит на мысль о том, что здесь значительно *меньше* случайности, чем можно было ожидать для этой модели. Какое техническое предположение для проверки гипотезы может быть здесь не выполнено?

21. Ракурс (угол расположения камеры) при съемке рекламного объекта может даже повлиять на оценку потребителем товара. В статье описан главный эффект угла расположения камеры при съемке ($F_{2,29} = 14,48$, $p < 0,001$), полученный с помощью дисперсионного анализа.[14] Средняя оценка для такого угла съемки равна 4,51, когда камера расположена на уровне глаз; 5,49 — для меньшего угла (выше уровня глаз) и 3,61 — для камеры, расположенной ниже уровня глаз. Более высокие баллы представляют более положительную оценку товара (персональный компьютер). Значимо ли различаются эти оценки в зависимости от угла съемки? Если да, то какой угол является наилучшим для съемок рекламы?

[14] Meyers-Levy Joan and Perrachio Laura A. "Getting an Angle in Advertising: The Effect of Camera Angle on Product Evaluations", *Journal of Marketing Research*, 29, p. 454–461.

22. Другой эксперимент, описанный теми же авторами, Мейерс-Леви и Перачио (Meyers-Levy and Peracchio), включал оценку изображения велосипеда, снятого под разными углами. Оценку давали две группы с различными уровнями мотивации. (Группа с высоким уровнем мотивации полагала, что она имеет хороший шанс выиграть велосипед.) Средние значения оценок были выше, если велосипед был снят при расположении камеры ниже или на уровне глаз, и значения оценок были ниже при угле камеры, расположенной выше уровня глаз (если на велосипед смотреть сверху вниз). Эти расхождения в оценках были выше для группы с низкой мотивацией. Результаты дисперсионного анализа оценки велосипеда включали изучение главного эффекта угла расположения камеры при съемке ($F_{2,106} = 7,00$, $p < 0,001$), главного эффекта мотивации ($F_{1,106} = 3,78$, $p < 0,05$) и эффекта их взаимодействия ($F_{2,106} = 3,83$, $p < 0,03$).

а) Значимо ли различаются средние оценки в группах с высокой и с низкой мотивацией? Обоснуйте свой ответ.

б) Можно ли на основе представленной здесь информации о дисперсионном анализе определить, какая из групп (группа с высокой или с низкой мотивацией) дала в среднем более высокие оценки?

в) Значимо ли взаимодействие угла расположения камеры при съемке и мотивации группы? Обоснуйте свой ответ.

г) Можно ли сделать заключение, что угол расположения камеры при съемке сильнее влияет на оценки группы с низкой мотивацией, чем с высокой, или влияние угла расположения камеры при съемке на оценки обеих групп одинаково, за исключением случайности? Объясните свой ответ.

Упражнения с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении A.

1. Объедините размеры годовой заработной платы в три группы в соответствии с уровнем квалификации служащих (A, B и C).

а) Для сравнения этих трех групп постройте блочные диаграммы и прокомментируйте их.

б) Найдите среднее значение для каждого уровня квалификации и прокомментируйте его.

в) Найдите межгрупповую и внутригрупповую вариацию и соответствующие им степени свободы.

г) Найдите значение F-статистики и количество степеней свободы для нее.

д) Выполните F-тест на уровне 0,05 и представьте результаты.

е) Изложите результаты теста наименьшего значимого различия, если он может быть применен.

ж) Обобщите, что вы узнали из базы данных относительно этой проблемы.

2. Ответьте на вопросы упражнения 1, заменив размер годовой заработной платы на возраст служащих.
3. Ответьте на вопросы упражнения 1, заменив размер годовой заработной платы на стаж работы служащих.

Проекты



1. Выберите некоторый количественный показатель, представляющий для вас интерес, и найдите его значение в Internet или в библиотеке хотя бы для 10 фирм в каждой из по меньшей мере трех промышленных групп. Вы получите как минимум 30 чисел.

   а) Постройте блочные диаграммы, по одной для каждой промышленной группы, и рассмотрите в целом свой набор данных. Для облегчения расчетов используйте одну и ту же шкалу.

   б) Найдите среднее и стандартное отклонения для каждой промышленной группы.

   в) Объясните, почему предположения для выполнения однофакторного дисперсионного анализа для ваших данных (1) удовлетворяются, (2) частично удовлетворяются, (3) совершенно не удовлетворяются. Исправьте, если это возможно, все серьезные проблемы путем преобразования данных.

   г) Найдите межгрупповую вариацию и соответствующие ей степени свободы.

   д) Найдите внутригрупповую вариацию и соответствующие ей степени свободы.

   е) Найдите значение F-статистики и количество степеней свободы для нее.

   ж) Найдите в F-таблице соответствующее критическое значение для выбранного вами уровня значимости.

   з) Выполните F-тест и изложите результаты.

   и) Если F-тест демонстрирует значимость, выполните тест наименьшего значимого различия для каждой пары промышленных групп и обобщите полученные результаты в отношении найденных различий.

2. Выберите в библиотеке несколько журналов из интересующей вас области бизнеса. Просмотрите статьи по нескольким проблемам, чтобы выбрать одну, где используется дисперсионный анализ. Напишите страницу текста, изложив следующее.

   а) Какой главный вопрос рассматривается?

   б) Какие данные проанализированы? Как они были получены?

   в) Найдите выполненную в статье проверку гипотезы. Определите нулевую и альтернативную гипотезы. Найдите результаты проверки.

ГЛАВА **16**

Непараметрические методы: проверка гипотез для порядковых данных или данных, не подчиняющихся нормальному распределению



Беспокоит ли вас то, что статистический вывод требует выполнения определенных допущений? Вероятно, да. В частности, может беспокоить то, что генеральная совокупность должна подчиняться *нормальному* распределению, а это так трудно проверить на основании выборочных данных. Конечно, иногда помогает центральная предельная теорема, но если размер выборки недостаточно велик, или распределение значительно скошено, или данные содержат сильно отличающиеся значения, было бы неплохо иметь альтернативу. И она есть.

Непараметрические методы представляют собой такие статистические процедуры проверки гипотез, которые не требуют нормальности (или любой другой формы) распределения данных, поскольку используют частоты или ранги (наименьшее наблюдаемое значение имеет ранг 1, следующее — 2, затем — 3 и т.д.) вместо фактических значений данных. Эти методы по-прежнему требуют наличия случайной выборки из генеральной совокупности, что обеспечивает наличие в данных необходимой информации. Поскольку эти методы основаны на ранжировании (упорядочении значений) и не используют сумм значений, многие непараметрические методы работают не только с количествен-

ными, но и с *порядковыми* данными. Ниже приведены формулировки двух параметрических подходов.

Непараметрический подход, основанный на частотах

1. Подсчитайте, сколько раз данное событие встречается в наборе данных.
2. Используйте биномиальное распределение, чтобы решить, согласуется ли полученная частота с нулевой гипотезой.

Непараметрический подход, основанный на рангах

1. Создайте новый набор данных, состоящий из рангов значений. **Ранг** значения показывает позицию этого значения после упорядочения всех данных. Например, набор данных (35, 95, 48, 38, 57) преобразуется в (1, 5, 3, 2, 4), так как значение 35 является наименьшим (оно имеет ранг 1), значение 95 — наибольшее (с рангом 5), значение 48 — третье наименьшее (ранг 3) и т.д.
2. Далее работайте не с исходными данными, а с рангами.
3. Используйте статистические формулы и таблицы, разработанные специально для проверки гипотез о рангах.

Параметрические методы представляют собой статистические процедуры, которые требуют полностью определенной модели. Большая часть рассмотренных нами процедур статистического вывода требовала параметрических моделей (включая t-тест, тесты для регрессий и F-тест). Например, линейная модель для регрессии определяет как уравнение прогноза, так и точную форму для случайного шума. В отличие от параметрических, непараметрические методы являются более гибкими и не требуют точного определения ситуации.

Непараметрические методы содержат одну большую и приятную неожиданность: вы теряете немного от того, что не используете преимущества нормального распределения (когда распределение действительно нормальное), но выигрываете очень много, если распределение действительно не является нормальным. Таким образом, использование непараметрического метода похоже на покупку страхового полиса: вы платите небольшие издержки, но если возникают проблемы, вы получаете солидную компенсацию.

Одним из способов оценки результативности различных статистических методов является сравнение их в отношении эффективности. Говорят, что некоторый статистический тест **эффективнее** другого, если он лучше использует информацию, содержащуюся в данных[1]. Таким образом, непараметрические методы почти так же эффективны, как параметрические в случае нормального распределения, и намного эффективнее при его отсутствии. Ниже приведены преимущества непараметрического подхода.

Преимущества непараметрических методов

1. Не нужны предположения о нормальности; могут быть использованы, даже если распределение не является нормальным.
2. Отсутствуют многие проблемы преобразования данных; можно использовать, даже если данные не могут быть просто преобразованы в форму с нормальным распределением, и фактически дают тот же результат, независимо от того, преобразовывались данные или нет.
3. Можно использовать для тестирования порядковых данных, так как ранги могут быть установлены исходя из естественного упорядочения.
4. Могут быть намного эффективнее параметрических методов в случае, когда распределение данных не является нормальным.

Существует только один, и то относительно небольшой, недостаток непараметрических методов.

Недостаток непараметрических методов

Менее статистически эффективен, чем параметрические методы, в случае нормального распределения; однако часто это снижение эффективности незначительно.

В этой главе мы проведем проверку гипотезы для одной выборки (проверку в отношении медианы), а также проверки гипотез двух независимых и для двух связанных выборок (проверку в отношении разности).

16.1. Проверка гипотезы о равенстве медианы некоторому заданному значению

С одной стороны, имея обычную выборку значений одной переменной из генеральной совокупности, можно использовать среднее и стандартную ошибку для проверки гипотезы относительно среднего генеральной совокупности (t-тест). И этот метод хорошо работает в случае нормального распределения значений этой переменной.

С другой стороны, непараметрический подход, поскольку он основан на использовании рангов упорядоченных данных, позволяет осуществить проверку гипотезы о *медиане* генеральной совокупности. Медиана является подходящим для данного случая показателем, поскольку определяется в терминах рангов. (Напомним, что медиана имеет ранг $(1+n)/2$ для выборки размером n.)

Как мы в данном случае можем освободиться от предположения о нормальности распределения? Это довольно легко объяснить: если распределение совокупности является непрерывным, то половина значений совокупности лежит выше медианы, а половина ниже[2]. Ввиду того что набор данных представляет собой случайную выборку независимых наблюдений, для него характерно биномиальное распределение вероятности. Используя терминологию из глав 6 и 7, можно сказать, что количество значений ниже медианы генеральной совокупности является количеством событий "попадания ниже медианы", которые произошли в результате n независимых испытаний при условии, что вероятность каждого события равна 1/2. В результате можно сделать следующий вывод.

[1] Формальное определение эффективности дается в терминах относительной работы (размеры выборок), необходимой для каждого из тестов, чтобы получить аналогичные результаты.

[2] Далее будет показано, как этот же тест может быть использован и для дискретной генеральной совокупности.

Количество значений в выборке, расположенных ниже медианы непрерывной генеральной совокупности, подчиняется биномиальному распределению с параметрами $p = 0,5$ и n, равном размеру выборки.

Критерий знаков

Критерий знаков использует это биномиальное распределение. Чтобы проверить, может ли медиана генеральной совокупности быть равна, например, $65 536, следует выяснить, сколько выборка содержит значений, которые меньше 65 536, и достоверность этого события исходя из биномиального распределения. **Критерий знаков** позволяет принять решение о том, равна ли медиана в генеральной совокупности некоторому заданному значению исходя из количества таких значений в выборке, которые находятся ниже этого заданного значения. К значениям не применяются никакие арифметические операции — только сравнение и подсчет. Эта процедура имеет следующий вид.

Критерий знаков

1. Подсчитываем количество данных, значения которых *отличаются* от заданного опорного значения θ_0. Это число обозначим m и назовем **размером модифицированной выборки**.
2. Используя таблицу, находим границы для этого размера модифицированной выборки.
3. Подсчитываем количество значений, которые лежат ниже заданного опорного значения θ_0, и сравниваем это число с границами, найденными в таблице[3].
4. Если найденное на шаге 3 число находится *вне* границ, то различие является статистически значимым. Если это число совпадает с одной из границ или находится *внутри* границ, то различие не является статистически значимым.

Гипотезы

Сначала предположим, что распределение совокупности непрерывно. Нулевая гипотеза для критерия знаков утверждает, что медиана генеральной совокупности θ точно равна известной опорной величине θ_0. (Обычно предполагают, что значение опорной величины задано заранее, а не рассчитано на основе рассматриваемого набора данных.) Альтернативная (исследовательская) гипотеза утверждает противоположное: медиана генеральной совокупности не равна опорной величине.

Гипотезы для критерия знаков в отношении медианы распределения непрерывной генеральной совокупности

$$H_0: \theta = \theta_0;$$

$$H_1: \theta \neq \theta_0,$$

где θ — неизвестное значение медианы генеральной совокупности, а θ_0 — известное значение опорной величины.

[3] Можно подсчитывать количество значений, лежащих *выше* заданного опорного значения. Результат проверки будет тем же.

В общем, даже если распределение не является непрерывным, критерий знаков установит, делит ли опорная величина θ_0 совокупность точно пополам[4].

Обобщенная формулировка гипотез для критерия знаков

H_0: в генеральной совокупности вероятность того, что значение превышает θ_0, равна вероятности того, что значение ниже, чем θ_0;

H_1: эти вероятности не равны,

где θ_0 — известное опорное значение, которое тестируется.

Допущение

Есть только одно допущение, выполнение которого необходимо для применения критерия знаков. Одним из преимуществ непараметрических методов является то, что для их использования нужно выполнить так немного требований.

Допущение, необходимое для критерия знаков

Рассматриваемый набор данных должен быть случайной выборкой из изучаемой генеральной совокупности.

В табл. 16.1 содержатся интервалы для критерия знаков.

Таблица 16.1. Интервалы для критерия знаков

Размер модифицированной выборки, m	Уровень значимости 5%			Уровень значимости 1%		
	Критерий знаков статистически значим, если число либо			Критерий знаков статистически значим, если число либо		
	меньше, чем	либо	больше, чем	меньше, чем	либо	больше, чем
6	1		5	–		–
7	1		6	–		–
8	1		7	1		7
9	2		7	1		8
10	2		8	1		9
11	2		9	1		10
12	3		9	2		10
13	3		10	2		11
14	3		11	2		12
15	4		11	3		12
16	4		12	3		13

[4] Это несколько отличается от того, является ли θ_0 медианой. Например, очень маленькая совокупность, состоящая из чисел (11, 12, 13, 13, 14), имеет медиану, равную 13. Однако два значения лежат ниже медианы, а одно выше. Таким образом, нулевая гипотеза для критерия знаков утверждает *больше*, чем только то, что медиана генеральной совокупности равна θ_0.

Продолжение табл. 16.1

Размер модифицированной выборки, m	Уровень значимости 5%			Уровень значимости 1%		
	Критерий знаков статистически значим, если число либо			Критерий знаков статистически значим, если число либо		
	меньше, чем	либо	больше, чем	меньше, чем	либо	больше, чем
17	5		12	3		14
18	5		13	4		14
19	5		14	4		15
20	6		14	4		16
21	6		15	5		16
22	6		16	5		17
23	7		16	5		18
24	7		17	6		18
25	8		17	6		19
26	8		18	7		19
27	8		19	7		20
28	9		19	7		21
29	9		20	8		21
30	10		20	8		22
31	10		21	8		23
32	10		22	9		23
33	11		22	9		24
34	11		23	10		24
35	12		23	10		25
36	12		24	10		26
37	13		24	11		26
38	13		25	11		27
39	13		26	12		27
40	14		26	12		28
41	14		27	12		29
42	15		27	13		29
43	15		28	13		30
44	16		28	14		30
45	16		29	14		31
46	16		30	14		32
47	17		30	15		32
48	17		31	15		33
49	18		31	16		33
50	18		32	16		34

Продолжение табл. 16.1

Размер модифицированной выборки, m	Уровень значимости 5%			Уровень значимости 1%		
	Критерий знаков статистически значим, если число либо			Критерий знаков статистически значим, если число либо		
	меньше, чем	либо	больше, чем	меньше, чем	либо	больше, чем
51	19		32	16		35
52	19		33	17		35
53	19		34	17		36
54	20		34	18		36
55	20		35	18		37
56	21		35	18		38
57	21		36	19		38
58	22		36	19		39
59	22		37	20		39
60	22		38	20		40
61	23		38	21		40
62	23		39	21		41
63	24		39	21		42
64	24		40	22		42
65	25		40	22		43
66	25		41	23		43
67	26		41	23		44
68	26		42	23		45
69	26		43	24		45
70	27		43	24		46
71	27		44	25		46
72	28		44	25		47
73	28		45	26		47
74	29		45	26		48
75	29		46	26		49
76	29		47	27		49
77	30		47	27		50
78	30		48	28		50
79	31		48	28		51
80	31		49	29		51
81	32		49	29		52
82	32		50	29		53
83	33		50	30		53

Окончание табл. 16.1

Размер модифицированной выборки, m	Уровень значимости 5%		Уровень значимости 1%	
	Критерий знаков статистически значим, если число либо		Критерий знаков статистически значим, если число либо	
	меньше, чем либо	больше, чем	меньше, чем либо	больше, чем
84	33	51	30	54
85	33	52	31	54
86	34	52	31	55
87	34	53	32	55
88	35	53	32	56
89	35	54	32	57
90	36	54	33	57
91	36	55	33	58
92	37	55	34	58
93	37	56	34	59
94	38	56	35	59
95	38	57	35	60
96	38	58	35	61
97	39	58	36	61
98	39	59	36	62
99	40	59	37	62
100	40	60	37	63

Если m больше 100, то для уровня значимости 0,05 табличные значения находят округлением $\left(m-1,96\sqrt{m}\right)/2$ и $\left(m+1,96\sqrt{m}\right)/2$ до ближайших целых чисел. Например, для m = 120 по формуле получаем значения 49,3 и 70,7 и округляем их до табличных значений 49 и 71 соответственно. При уровне значимости, равном 0,01, округляют значения, полученные по формулам $\left(m-2,576\sqrt{m}\right)/2$ и $\left(m+2,576\sqrt{m}\right)/2$.

Пример. *Сравнение доходов семьи на местном и национальном уровнях*

Вы собираетесь открыть высококлассные фирменные рестораны в нескольких населенных пунктах. В каждом из этих населенных пунктов вас интересует медиана семейного дохода, поскольку среднее значение дохода может повышаться за счет лишь небольшого количества семей. Проведенное в одном из населенных пунктов исследование дало оценку медианы семейного дохода $70 547, и вас интересует, значимо ли это значение выше медианы семейного дохода в целом по стране, равного $27 735[5]. Кажется, что в этом населенном пункте значение медианы дохода выше, но не следует спешить с таким выводом на основании выборки из 25 семей. В табл. 16.1.2 приведен набор данных и отмечены те семьи, которые имеют доход ниже $27 735.

[5] Из отчета Bureau of the Census за 1985 год, *Statistical Abstracts of the United States, 1987* (Washington D. C., 1986), p. 437.

Таблица 16.1.2. Доходы семей, которые попали в выборку (в долларах)

39 465	96 270	16 477*	138 933
80 806	85 421	5 921*	70 547
267 525	56 240	187 445	81 802
163 819	14 706*	83 414	78 464
58 525	54 348	36 346	
25 479*	7 081*	19 605*	
29 341	137 414	156 681	

* Доход ниже $27 735.

Значение опорной величины $\theta_0 = \$27735$, причем это значение взято не из рассматриваемого набора данных. Ниже приведена последовательность выполнения критерия знаков.

1. Все 25 семей имеют значения дохода, отличные от значения опорной величины, так что размер модифицированной выборки m = 25, т.е. равен размеру исходной выборки.
2. Для m = 25 и уровня значимости 5% находим в таблице значения границ 8 и 17.
3. Выборка содержит шесть семей с доходом ниже значения опорной величины.
4. Ввиду того что число 6 выходит за границы (поскольку оно меньше 8), нулевую гипотезу отвергают и делают вывод, что результат является статистически значимым.

Значение медианы семейного дохода для данного населенного пункта, $70 547, значимо отличается от значения медианы семейного дохода в целом по стране — $27 735.

Ваши предположения подтвердились: это населенный пункт с высокими доходами жителей. Медиана дохода семьи в данном населенном пункте значимо выше медианы дохода семьи по стране в целом[6].

16.2. Тестирование различий в двух связанных выборках

Если набор данных состоит из *пар* наблюдений, размещенных в двух столбиках, то можно создать одну выборку, представляющую собой изменения или разности между элементами пар. Такие данные могут быть получены в исследованиях типа "до/после", где измерения выполняются до и после некоторого вмешательства (демонстрация рекламы, лечение, регулировка приборов и т.п.). Как выполнять t-тест для двух связанных (зависимых) выборок, описано в главе 10. Здесь же мы рассмотрим непараметрическое решение.

Использование критерия знаков для разностей

Непараметрическая процедура проверки того, значимо ли отличаются значения в двух колонках (**критерий знаков для разностей**), заключается в применении описанного в предыдущем разделе критерия знаков к данным одной колонки, содержащей разности значений двух исходных колонок. Значение опорной величины θ_0 в этом случае равно 0, что соответствует отсутствию различий в генеральной совокупности. Критерий знаков показывает, являются ли эти изменения сбаланси-

[6] Это одностороннее заключение для двустороннего теста, как описано в главе 10.

рованными (т.е. количество увеличений такое же, как и количество уменьшений, за исключением случайности), или имеются систематические различия (например, количество увеличений значимо больше количества уменьшений).

В табл. 16.2.1 показано, как выглядит в таком случае типичный набор данных.

Таблица 16.2.1. Парные (связанные) наблюдения

Единица наблюдения	Колонка 1	Колонка 2
1	X_1	Y_1
2	X_2	Y_2
.	.	.
.	.	.
.	.	.
n	X_n	Y_n

Обычно колонка 1 (X) содержит результат измерения "до", а колонка 2 (Y) — "после". Важно, чтобы была естественная взаимосвязь этих двух колонок, чтобы каждая строка содержала результаты двух наблюдений (измеренные в одних и тех же единицах), относящиеся к *одному и тому же* объекту. Ниже приведена процедура проверки.

Критерий знаков для разностей

1. Подсчитайте количество различных значений разности (изменений) между колонками 1 и 2. Это и есть размер модифицированной выборки m.
2. По таблице найдите границы для такого размера модифицированной выборки.
3. Подсчитайте количество объектов, у которых произошло уменьшение (т.е. у них значение в колонке 2 оказалось меньше значения в колонке 1), и сравните это число со значениями границ, найденными по таблице[7].
4. Если полученное на шаге 3 число находится *вне* границ, то две выборки значимо различаются. Если это число *совпадает* с одной из границ или находится *внутри* границ, то между двумя выборками нет значимого различия.

Обратите внимание, что имеет значение только направление изменения значения между колонками (возрастание или уменьшение), а не фактическая величина изменения. Это означает, что этот тест можно использовать и для порядковых, и для количественных данных. Главное, чтобы существовало некоторое упорядочение, позволяющее определить направление изменения.

Гипотезы

Нулевая гипотеза утверждает, что в генеральной совокупности количество изменений в большую сторону равно количеству изменений в меньшую сторону (при сравнении связанных пар значений X и Y). В соответствии с этой гипотезой изменения в сторону увеличения или уменьшения происходят просто случайно. Исследовательская (альтернативная) гипотеза утверждает, что вероятности изменений в большую и в меньшую сторону различаются между собой.

[7] Вместо этого можно подсчитывать количество объектов, для которых произошло *увеличение*. Результат тестирования от этого не изменится.

Гипотезы для критерия знаков в отношении разностей

H_0: вероятность того, что $X < Y$, равна вероятности того, что $Y < X$. Другими словами, вероятность увеличения значения равна вероятности уменьшения.

H_1: вероятность того, что $X < Y$ *не* равна вероятности того, что $Y < X$. Иначе говоря, вероятности увеличения и уменьшения значений не равны между собой.

Условие

Как и для других подобных тестов, для возможности применять критерий знаков к разностям необходимым является выполнение следующего условия.

Условие применимости критерия знаков для разностей

Набор данных представляет собой случайную выборку из рассматриваемой генеральной совокупности. Каждой элементарной единице этой совокупности соответствуют два значения, X и Y, измеренных в одних единицах.

Пример. *Оценка двух реклам*

Каждому члену группы из 17 человек показали две рекламы. Каждый из испытуемых оценил творческий уровень каждой из реклам. Результаты приведены в табл. 16.2.2.

Количество различающихся между собой оценок двух реклам равно 13. Другими словами, 13 человек оценили эти две рекламы по-разному, а четверо — одинаково. Таким образом, размер модифицированной выборки равен 13.

Найденные по таблице границы для уровня значимости 5% и размера выборки $m = 13$ равны 3 и 10.

Три человека оценили рекламу 2 ниже рекламы 1. Это число находится внутри границ (чтобы результат был значимым, число людей должно быть *либо меньше 3, либо больше 10*).

Поэтому принимаем нулевую гипотезу, которая утверждает, что творческие уровни реклам одинаковы.

Результат не является статистически значимым. Несмотря на то что 3 человека из 17 оценили выше рекламу 1, это может быть обусловлено случайностью, а не особыми качествами реклам.

Таблица 16.2.2. Творческий уровень рекламы

Реклама 1	Реклама 2	Реклама 1	Реклама 2
4	2	5	4
2	4	3	4
4	5	3	5
4	4	4	5
4	4	5	5
2	5	4	5
3	3	5	4
4	5	2	5
3	5		

16.3. Проверка значимости различия двух независимых выборок

Предположим, что мы имеем две независимые (несвязанные) выборки и хотим узнать, взяты ли они из генеральных совокупностей с одинаковым распределением или нет. С одной стороны, описанный в главе 10 t-тест для независимых выборок предполагает, что генеральные совокупности распределены нормально (с одинаковым стандартным отклонением для небольших выборок), и затем проверяет равенство средних этих генеральных совокупностей. С другой стороны, непараметрический метод предполагает лишь то, что есть две случайные выборки из двух генеральных совокупностей, и затем проверяет, одинаково ли распределены эти генеральные совокупности. В табл. 16.3.1 содержится типичный набор данных для двух несвязанных выборок.

Процедура, основанная на ранжировании *всех* данных

Для выполнения теста необходимо, во-первых, *объединить данные обеих выборок* и вычислить общие ранги значений в полученном наборе данных. Если значения в одной выборке систематически меньше значений другой выборки, то и ранги этих значений будут соответственно меньше. Сравнивая ранги значений одной выборки с рангами значений другой, можно выяснить, различаются ли они систематически или просто случайно.

Существует несколько способов получить ответ на этот вопрос. **Критерий суммы рангов Вилкоксона** и **U-критерий Манна–Уитни** — это два разных способа получения одного и того же результата относительно непараметрического теста для двух независимых выборок. Критерий суммы рангов Вилкоксона использует сумму общих рангов одной из выборок, а в основе U-критерия Манна–Уитни лежит количество способов, с помощью которых в одной выборке можно найти значение, превышающее значение в другой выборке.

Проще работать со *средним рангом* двух выборок. Этот критерий алгебраически эквивалентен другим (в том смысле, что критерий Вилкоксона, критерий Манна–Уитни, а также приведенный здесь критерий разности средних рангов дают один и тот же результат) и ясно показывает, что, несмотря на то, что в непараметрических методах оперируют рангами, а не значениями данных, здесь используются те же основные идеи статистики[8]. Ниже приведена процедура выполнения теста.

Таблица 16.3.1. Две несвязанные выборки

Выборка 1 (n_1 наблюдений из совокупности 1)	Выборка 2 (n_2 наблюдений из совокупности 2)
$X_{1,1}$	$X_{2,1}$
$X_{1,2}$	$X_{2,2}$
.	.
.	.
.	.
X_{1,n_1}	X_{2,n_2}

Примечание. Размеры выборок n_1 и n_2 могут быть разными.

Непараметрический критерий для двух независимых выборок

1. Объедините данные обеих выборок вместе и упорядочьте их (т.е. расположите значения в порядке возрастания) для получения общих рангов (т.е. рангов значений в объединенном наборе). Если есть одинаковые значения, то им присваивается одинаковый ранг, равный среднему значению рангов этих значений, чтобы одинаковые значения имели одинаковые ранги.
2. Найдите среднее значение всех рангов для каждой выборки, $\overline{R_1}$ и $\overline{R_2}$.
3. Определите разность между этими средними, $\overline{R_2} - \overline{R_1}$.
4. Вычислите стандартную ошибку для разности средних значений рангов[9]:
5. $(n_1 + n_2)\sqrt{\dfrac{n_1 + n_2 + 1}{12 n_1 n_2}}$.
6. Вычислите значение тест-статистики, разделив разность средних (полученную на шаге 3) на значение стандартной ошибки (полученное на шаге 4):
7. $$\text{Тест-статистика} = \frac{\overline{R_2} - \overline{R_1}}{(n_1 + n_2)\sqrt{\dfrac{n_1 + n_2 + 1}{12 n_1 n_2}}}.$$
8. Если значение тест-статистики превышает 1,960, то две выборки *значимо различаются*. Если значение тест-статистики меньше 1,960, то две выборки *не имеют значимых различий*[10].

Гипотезы

Нулевая гипотеза утверждает, что две выборки были извлечены из генеральных совокупностей с *одинаковым распределением*. Исследуемая (альтернативная) гипотеза утверждает, что соответствующие генеральные совокупности имеют *разное распределение*.

Гипотезы тестирования двух независимых выборок

H_0 : две выборки извлечены из генеральных совокупностей с одинаковым распределением.

H_1 : две выборки извлечены из совокупностей, имеющих разные распределения.

[8] Например, формула, выражающая разность средних значений рангов в терминах U-статистики, имеет вид $(n_1 + n_2)(U - n_1 n_2 / 2)/(n_1 n_2)$. U-статистику Манна–Уитни определяют как $n_1 n_2 + n_1(n_1 + 1)/2$ минус сумма общих рангов одной из выборок.

[9] Стандартная ошибка имеет точное значение при отсутствии в выборках одинаковых значений. Здесь нет необходимости строить оценку, поскольку можно произвести непосредственные вычисления исходя из свойств случайно перемешанных (в соответствии с нулевой гипотезой) рангов.

[10] Чтобы использовать другой уровень значимости, вместо числа 1,960 следует взять из t-таблицы соответствующее значение для бесконечного числа степеней свободы. Например, для уровня значимости 1% вместо 1,960 следует использовать число 2,576.

Допущения

Для того чтобы применять тест для двух независимых выборок, необходимо выполнение следующих допущений. В дополнение к обычному требованию о случайности выборок, для того, чтобы можно было использовать из t-таблицы значения для бесконечного количества степеней свободы, необходимо, чтобы размеры выборок были достаточно большими.

Допущения, необходимые для теста двух независимых выборок

1. Каждая выборка является случайной выборкой из соответствующей генеральной совокупности.
2. Из каждой генеральной совокупности выбрано больше 10 элементов, т.е. $n_1 > 10$ и $n_2 > 10$.

Пример. *Ссуда под недвижимость с фиксированной и регулируемой процентными ставками*

Банк планирует проведение маркетинговой кампании для выдачи ссуды под залог недвижимости. На собрании некоторые сотрудники высказали мнение, что изменяемая процентная ставка ссуды под залог недвижимости более привлекательна для заемщиков с низкими доходами, потому что они смогут получить большую ссуду и смогут позволить себе купить более дорогой дом. Другие предположили, что повышенный риск изменяемой процентной ставки ссуды ориентирован больше на заемщиков с высокими доходами, потому что у них есть "амортизатор" на случай, если в будущем суммы их платежей пойдут вверх. Какая из этих двух групп права? Вы собрали данные о доходах тех, кто обращался за ссудой под залог недвижимости. Эти данные приведены в табл. 16.3.2.

Обратите внимание на наличие одного сильно отличающегося значения дохода ($240 000). Это одно

Таблица 16.3.2. Размеры доходов тех, кто обращался за ссудой под залог недвижимости

Фиксированная процентная ставка, дол.	Изменяемая процентная ставка, дол.
34 000	37 500
25 000	86 500
41 000	36 500
57 000	65 500
79 000	21 500
22 500	36 500
30 000	99 500
17 000	36 000
36 500	91 000
28 000	59 500
240 000	31 000
22 000	88 000
57 000	35 500
68 000	72 000
58 000	
49 500	

значение вызывает проблему при использовании t-теста для двух выборок. Вы проверили и выяснили, что это корректное значение. Можно ли исключить это значение из анализа? Вероятно, нет, поскольку оно представляет семью с высоким доходом, которая просит ссуду с фиксированным процентом, и, следовательно, это полезная информация.

Спасением является использование непараметрического метода! Такие методы оперируют рангами, а не значениями данных, и поэтому величина дохода не имеет значения, даже если она равна одному миллиарду долларов. Такое значение будет рассматриваться просто как самое большое.

На рис. 16.3.1 данные двух выборок представлены в виде блочных диаграмм. Рисунок свидетельствует, что в группе с изменяемой процентной ставкой доходы более высокие, но это не совсем очевидно из-за значительного пересечения диапазонов значений в этих двух группах.

Необходимо упорядочить все значения доходов. Для этого создаем новый набор данных иной структуры, содержащий одну колонку значений доходов (куда записывают данные обеих колонок исходного набора данных) и другую колонку с видом ссуды. Полученный набор данных показан в табл. 16.3.3.

Теперь вы готовы упорядочить (ранжировать) все строки базы данных в соответствии со значением дохода. Результаты ранжирования показаны в табл. 16.3.4. С помощью электронной таблицы можно также провести ранжирование совместно по доходу и типу ссуды. После ранжирования каждый ранг обозначен числом 1, 2, 3 и т.д.

Если имеются совпадения (два или больше одинаковых значения доходов), то для этих значений используют среднее значение ранга. Например, в табл. 16.3.4 доход в $36 500 встречается трижды (при рангах 12, 13 и 14), так что среднее значение ранга для значения дохода, равного $36 500, будет равно (12 + 13 + 14)/3 =13. Два дохода равны $57 000, и среднее значение ранга равно для них (18 + 19)/2 = 18,5.

Теперь можно вычислить среднее рангов. Из колонки рангов сначала выбирают только числа, представляющие ссуду с фиксированной процентной ставкой, и вычисляют среднее значение ранга:

$$\frac{1+3+4+5+6+7+9+13+16+17+18{,}5+18{,}5+20+23+25+30}{16}=\frac{216}{16}=13{,}50\,.$$

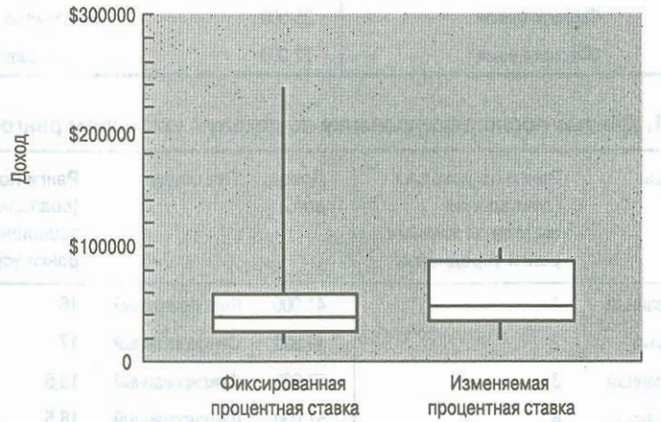


Рис. 16.3.1. Несмотря на то что самое большое значение дохода находится в группе с фиксированной процентной ставкой, доходы в группе с изменяемой процентной ставкой в целом фактически выше. Однако имеет место значительное пересечение диапазонов значений доходов двух групп. Наличие очень большого, резко отличающегося от всех остальных, значения дохода является препятствием для t-теста для двух выборок, но не является препятствием для непараметрического теста

Далее вычисляют среднее значение ранга для ссуды с изменяемой процентной ставкой:

$$\frac{2+8+10+11+13+13+15+21+22+24+26+27+28+29}{14}=\frac{249}{14}=17{,}7857\,.$$

В табл. 16.3.5 показан полученный результат.

Таблица 16.3.3. Начальные данные перед ранжированием

Доход, дол.	Тип ссуды	Доход, дол.	Тип ссуды
34 000	Фиксированная	49 500	Фиксированная
25 000	Фиксированная	37 500	Изменяемая
41 000	Фиксированная	86 500	Изменяемая
57 000	Фиксированная	36 500	Изменяемая
79 000	Фиксированная	65 500	Изменяемая
22 500	Фиксированная	21 500	Изменяемая
30 000	Фиксированная	36 500	Изменяемая
17 000	Фиксированная	99 500	Изменяемая
36 500	Фиксированная	36 000	Изменяемая
28 000	Фиксированная	91 000	Изменяемая
240 000	Фиксированная	59 500	Изменяемая
22 000	Фиксированная	31 000	Изменяемая
57 000	Фиксированная	88 000	Изменяемая
68 000	Фиксированная	35 500	Изменяемая
58 000	Фиксированная	72 000	Изменяемая

Таблица 16.3.4. Данные после ранжирования по доходу с указанием рангов

Доход, дол.	Тип ссуды	Ранги по доходам (совпадения выделены жирным, ранги усреднены)	Доход, дол.	Тип ссуды	Ранги по доходам (совпадения выделены жирным, ранги усреднены)
17 000	Фиксированный	1	41 000	Фиксированный	16
21 500	Изменяемый	2	49 500	Фиксированный	17
22 000	Фиксированный	3	57 000	Фиксированный	18,5
22 500	Фиксированный	4	57 000	Фиксированный	18,5
25 000	Фиксированный	5	58 000	Фиксированный	20
28 000	Фиксированный	6	59 500	Изменяемый	21
30 000	Фиксированный	7	65 500	Изменяемый	22
31 000	Изменяемый	8	68 000	Фиксированный	23
34 000	Фиксированный	9	72 000	Изменяемый	24
35 500	Изменяемый	10	79 000	Фиксированный	25

Доход, дол.	Тип ссуды	Ранги по доходам (совпадения выделены жирным, ранги усреднены)	Доход, дол.	Тип ссуды	Ранги по доходам (совпадения выделены жирным, ранги усреднены)
36 000	Изменяемый	11	86 500	Изменяемый	26
36 500	Фиксированный	13	88 000	Изменяемый	27
36 500	Изменяемый	13	91 000	Изменяемый	28
36 500	Изменяемый	13	99 500	Изменяемый	29
37 500	Изменяемый	15	240 000	Фиксированный	30

Таблица 16.3.5. Доходы тех, кто обращался за ссудой под залог недвижимости, и соответствующие ранги

Ссуда с фиксированной процентной ставкой		Ссуда с изменяемой процентной ставкой	
доход, дол.	ранг	доход, дол.	ранг
34 000	9	37 500	15
25 000	5	86 500	26
41 000	16	36 500	13
57 000	18,5	65 500	22
79 000	25	21 500	2
22 500	4	36 500	13
30 000	7	99 500	29
17 000	1	36 000	11
36 500	13	91 000	28
28 000	6	59 500	21
240 000	30	31 000	8
22 000	3	88 000	27
57 000	18,5	35 500	10
68 000	23	72 00	24
58 000	20		
49 500	17		
Среднее значение ранга	$\overline{R}_1 = 13{,}50$		$\overline{R}_2 = 17{,}7857$

Размеры выборок: $n_1 = 16$, $n_2 = 14$.

Размеры доходов перечислены в исходном порядке, но с указанием рангов. Обратите внимание, что резко отличающееся значение ($240 000) имеет наивысший ранг, равный 30, но этот *ранг* уже не является резко отличающимся значением. Заметим также, что (как и предполагалось на основании блочной диаграммы) явно более низкий уровень дохода, соответствующий ссуде с фиксированной процентной ставкой, имеет и более низкое среднее значение ранга.

Но действительно ли различаются доходы тех, кто обращается за ссудой с фиксированной процентной ставкой, и тех, кто обращается за ссудой с изменяемой процентной ставкой? Иными словами, является ли разность средних значений рангов (13,50 и 17,79) статистически значимой? Используя соответствующую формулу, вычисляем стандартную ошибку разности средних значений рангов:

$$\text{Стандартная ошибка} = (n_1+n_2)\sqrt{\frac{n_1+n_2+1}{12n_1n_2}} =$$

$$= (16+14)\sqrt{\frac{16+14+1}{12\times16\times14}} =$$

$$= (30)\sqrt{\frac{31}{2,688}} = (30)\times0,10739 = 3,2217.$$

Тест-статистику вычисляем путем деления разности средних значений рангов на стандартную ошибку:

$$\text{Тест статистика} = \frac{\text{Разность средних значений рангов, } \overline{R_2}-\overline{R_1}}{\text{Стандартная ошибка}} =$$

$$= \frac{17,7857-13,500}{3,2217} = \frac{4,2857}{3,2217} = 1,3303.$$

Поскольку величина тест-статистики (рассматривают только абсолютное значение, знак "минус", если он есть, игнорируют) меньше 1,960, то делаем вывод, что две выборки не имеют значимых различий. Наблюдаемое различие между доходами тех, кто обращается за ссудой с фиксированной процентной ставкой, и тех, кто обращается за ссудой с изменяемой процентной ставкой, не является статистически значимым.

16.4. Дополнительный материал

Резюме

Непараметрические методы представляют собой такие статистические процедуры для проверки гипотез, которые не требуют нормального распределения данных (или любого иного определенного распределения), поскольку эти методы основаны на частотах или рангах, а не на реальных числовых значениях. Многие непараметрические методы работают с порядковыми данными так же хорошо, как и с количественными. Чтобы использовать непараметрический подход, основанный на частотах, необходимо выполнить следующие действия.

1. Подсчитать, сколько раз данное событие встречается в наборе данных.
2. Использовать биномиальное распределение для принятия решения о том, согласуется или нет полученная частота с нулевой гипотезой.

Чтобы использовать непараметрический подход, основанный на рангах, необходимо придерживаться такой последовательности действий.

1. Создать новый набор данных, преобразовав каждое значение данных в ранг. **Ранг** значения показывает его позицию после упорядочения всего набора данных. Например, набор данных (35, 95, 48, 38, 57) преобразуется в ряд (1, 5, 3, 2, 4), поскольку значение 35 является наименьшим (его ранг равен 1), значение 95 — наибольшим (с рангом 5), значение 48 является третьим, начиная с наименьшего (ранг 3), и т.д.

2. Перейти от рассмотрения данных к рассмотрению рангов.
3. Использовать специальные статистические формулы и таблицы для тестирования рангов.

Параметрические методы представляют собой статистические процедуры, которые требуют полностью определенной модели. Таково большинство методов, рассмотренных ранее. Один из важных вопросов заключается в том, какова эффективность непараметрических тестов по сравнению с параметрическими. Говорят, что некоторый тест **эффективнее** другого, если он лучше использует информацию, содержащуюся в данных. Непараметрические тесты имеют ряд преимуществ.

1. Нет необходимости предполагать нормальное распределение данных.
2. Отсутствуют многие проблемы преобразования данных. Фактически результат теста не зависит от того, преобразовывались данные или нет.
3. Можно использовать тест даже для порядковых данных, поскольку ранги могут быть основаны на некотором естественном упорядочении.
4. Если распределение данных действительно отличается от нормального, такие тесты могут быть намного эффективнее параметрических методов.

Единственным недостатком непараметрического метода тестирования является то, что он менее статистически эффективен, чем параметрический метод, в случае, если данные распределены нормально. Однако часто эта потеря эффективности незначительна.

Критерий знаков исходя из количества выборочных значений, которые попали ниже значения опорной величины, решает, равно ли значение медианы генеральной совокупности значению опорной величины. Вместо предположения о нормальном распределении теоретическое обоснование критерия основано на том факте, что количество таких значений в выборке, которые находятся ниже медианы генеральной совокупности с непрерывным распределением, подчиняется биномиальному распределению, для которого $\pi = 0,5$ и n — размер выборки.

Для выполнения критерия знаков необходимо выполнить следующие действия.

1. Подсчитать количество значений, которые отличаются от значения опорной величины θ_0. Это число m представляет собой **размер модифицированной выборки.**
2. По таблице найти для этого m значения границ.
3. Подсчитать количество данных, значения которых лежат ниже значения опорной величины θ_0, и сравнить это число с границами, найденными по таблице.
4. Если найденное на шаге 3 число находится вне границ, то различие является статистически значимым. Если это число находится на одной из границ или внутри границ, то различие не является статистически значимым.

Гипотезы для критерия знаков в отношении медианы непрерывно распределенной генеральной совокупности записываются следующим образом:

$$H_0 : \theta = \theta_0;$$
$$H_1 : \theta \neq \theta_0,$$

где θ — (неизвестное) значение медианы совокупности, а θ_0 — (известное) значение опорной величины, относительно которого проводится тестирование.

В общем виде гипотезы для критерия знаков формулируются следующим образом:

- H_0: В генеральной совокупности вероятность того, что значение больше, чем θ_0, равна вероятности того, что значение меньше, чем θ_0;
- H_1: эти вероятности не равны,

где θ_0 — (известное) значение опорной величины, относительно которого проводится тестирование. Предполагается, что набор данных представляет собой случайную выборку из рассматриваемой генеральной совокупности.

При наличии связанных попарно наблюдений (например, данных типа "измерение до" и "измерение после") критерий знаков можно применить к изменениям или разностям значений в этих парах. Непараметрическая процедура проверки значимости различий значений в двух колонках называется **критерием знаков для разностей**. Для выполнения процедуры необходимо выполнить следующие действия.

1. Подсчитать количество таких значений данных, которые отличаются в колонках 1 и 2. Это число, m, является *размером модифицированной выборки*.
2. Для этого значения m найти в таблице соответствующие границы.
3. Подсчитать количество значений, которые уменьшились (т.е. таких, значение которых в колонке 2 меньше значения в колонке 1), и сравнить это количество с границами, найденными по таблице.
4. Если полученное на шаге 3 число находится *вне* границ, то две выборки значимо различаются. Если это число *совпадает* с одной из границ или находится *внутри* границ, то различие между двумя выборками не является статистически значимым.

Гипотезы для критерия знаков в отношении разностей формулируются следующим образом:

- H_0: вероятность того, что $X < Y$, равна вероятности того, что $Y < X$. Другими словами, вероятность увеличения равна вероятности уменьшения.
- H_1: вероятность того, что $X < Y$, *не* равна вероятности того, что $Y < X$. Иначе говоря, вероятности увеличения и уменьшения не равны между собой.

Предполагается, что набор данных представляет собой такую случайную выборку из рассматриваемой совокупности, где для каждой элементарной единицы есть оба значения X и Y, измеренные в одних единицах.

Существует непараметрическая процедура, которую можно применять вместо t-теста для проверки различий двух независимых (несвязанных) выборок.

Критерий суммы рангов Вилкоксона и U-критерий Манна–Уитни — два разных способа непараметрического тестирования двух независимых выборок, которые дают один и тот же результат. Критерий суммы рангов Вилкоксона основан на сумме общих рангов одной из выборок, а в основе U-критерия Манна–Уитни лежит количество способов, с помощью которых можно в одной выборке найти значение, превышающее значение в другой выборке. Более простой способ заключается в использовании *среднего ранга* двух выборок.

1. Объедините данные обеих выборок и расположите значения в порядке возрастания для получения *набора общих рангов*. Если есть совпадающие значения, то в каждой такой группе всем значениям присваивают одинаковый ранг, равный среднему значению рангов в этой группе.
2. Найдите среднее значение общих рангов для каждой выборки, $\overline{R_1}$ и $\overline{R_2}$.
3. Вычислите разность между этими средними общих рангов, $\overline{R_2} - \overline{R_1}$.
4. Вычислите стандартную ошибку для разности средних значений рангов:

$$(n_1 + n_2)\sqrt{\frac{n_1 + n_2 + 1}{12 n_1 n_2}}.$$

5. Вычислите значение тест-статистики, разделив полученную на шаге 3 разность средних на полученное на шаге 4 значение стандартной ошибки.

$$\text{Тест-статистика} = \frac{\overline{R_2} - \overline{R_1}}{(n_1 + n_2)\sqrt{\frac{n_1 + n_2 + 1}{12 n_1 n_2}}}.$$

6. Если абсолютное значение тест-статистики больше 1,960, то две выборки *значимо различаются*. Если значение тест-статистики меньше 1,960, то *нет значимого различия* между этими двумя выборками.

Гипотезы для тестирования двух независимых выборок формулируются следующим образом:

- H_0: две выборки извлечены из генеральных совокупностей, имеющих одинаковое распределение;
- H_1: две выборки извлечены из генеральных совокупностей с разными распределениями.

Предположения, которые должны выполняться для того, чтобы можно было применять тест для двух независимых выборок.

1. Каждая выборка представляет собой случайную выборку из соответствующей генеральной совокупности.
2. Из каждой генеральной совокупности извлечено более 10 элементарных единиц, т.е. $n_1 > 10$ и $n_2 > 10$.

Основные термины

- Непараметрический метод (nonparametric method), 847
- Ранг (rank), 848
- Параметрический метод (parametric method), 848
- Эффективность (efficient), 848
- Проверка знаков (sign test), 850
- Размер модифицированной выборки (modified sample size), 850
- Критерий знаков для разностей (sign test for differences), 855
- Критерий суммы рангов Вилкоксона (Wilcoxon rank-sum test), 858
- U-критерий Манна–Уитни (Mann-Whitney U test), 858

Контрольные вопросы

1. а) Что такое непараметрический статистический метод?

   б) Что подсчитывают, когда используют непараметрический подход, основанный на частотах? Какое распределение вероятности используют для принятия решений?

   в) Какая информация из набора данных не принимается во внимание при использовании непараметрического подхода, основанного на рангах? Что используют вместо этой информации?

2. а) Что такое параметрический статистический метод?

   б) Назовите некоторые параметрические методы, которые вы использовали раньше.

3. а) Назовите преимущества непараметрических тестов по сравнению с параметрическими методами, если такие преимущества есть.

   б) Приведите недостатки непараметрических тестов по сравнению с параметрическими методами, если такие недостатки есть. Насколько серьезны эти недостатки?

4. а) Как вы можете интерпретировать утверждение: "Один тест эффективнее другого"?

   б) Если генеральная совокупность распределена нормально, какой из двух тестов будет более эффективен: параметрический или непараметрический?

   в) Если распределение генеральной совокупности сильно отличается от нормального, какой из двух тестов будет более эффективен: параметрический или непараметрический?

5. а) Для какой меры центральной тенденции непрерывной генеральной совокупности применяется критерий знаков?

   б) На каком вероятностном распределении основан этот тест?

   в) Предположим, что генеральная совокупность является дискретной и значительная часть совокупности *равна* медиане этой совокупности. Как в таком случае меняются гипотезы для критерия знаков в сравнении с непрерывной генеральной совокупностью?

6. а) Можно ли использовать критерий знаков для количественных данных? Почему?

   б) Можно ли использовать критерий знаков для порядковых данных? Почему?

   в) Можно ли использовать критерий знаков для номинальных данных? Почему?

7. а) Какое предположение должно выполняться для того, чтобы можно было применять критерий знаков?

   б) Выполнение какого предположения не требуется для критерия знаков, но обязательно для того, чтобы можно было применять t-тест?

8. Укажите на сходства и различия между критерием знаков и t-тестом.

9. а) Для какого вида данных применяется критерий знаков в отношении разности?

   б) Какие гипотезы в таком случае проверяются?

   в) Какое предположение должно выполняться?

10. а) Можно ли для количественных данных использовать критерий знаков для разностей? Почему?

    б) Можно ли для порядковых данных использовать критерий знаков для разностей? Почему?

    в) Можно ли для номинальных данных использовать критерий знаков для разностей? Почему?

11. Укажите на сходства и различия между критерием знаков для разностей и t-тестом для двух связанных выборок.

12. а) Опишите набор данных, состоящий из двух независимых (несвязанных) выборок.

    б) Какие гипотезы обычно проверяют для таких данных?

13. а) Чем отличается (если отличается) критерий суммы рангов Вилкоксона от U-критерия Манна–Уитни?

    б) Какая взаимосвязь существует между критерием суммы рангов Вилкоксона, U-критерием Манна–Уитни и тестом, основанным на разности средних значений общих рангов в каждой из выборок?

14. а) Что следует предпринять, если при проверке гипотезы для двух независимых (несвязанных) выборок в одной из выборок имеется сильно отличающееся значение? Для каждого из двух случаев (очень большое или очень малое сильно отличающееся значение) скажите, каким будет ранг этого сильно отличающегося значения?

    б) Какой статистический метод (параметрический или непараметрический) более чувствителен к наличию в данных сильно отличающихся значений? Почему?

15. а) Можно ли непараметрический тест для двух независимых выборок использовать для количественных данных? Почему?

    б) Можно ли непараметрический тест для двух независимых выборок использовать для порядковых данных? Почему?

    в) Можно ли непараметрический тест для двух независимых выборок использовать для номинальных данных? Почему?

16. Укажите на сходства и различия между непараметрическим тестом для двух независимых выборок и t-тестом для независимых выборок.

Задачи

1. Какие методы, параметрические или непараметрические, предпочтительнее в каждой из приведенных ниже ситуаций. Обоснуйте свой выбор и укажите, насколько серьезная проблема могла бы возникнуть при использовании другого метода.

а) Набор данных состоит из оценок облигаций, причем облигация AAA имеет более высокий разряд, чем облигация AA, а разряд облигации AA выше, чем разряд облигации A, и т.д.

б) Набор данных состоит из значений объема прибыли как процента от продаж, причем одно из значений сильно отличается от других, так как одной из фирм предъявлен серьезный иск. Вы считаете, что это значение необходимо учесть, поскольку такого рода иски представляют собой один из видов рисков в данной промышленной группе.

в) Набор данных состоит из значений весов вентилей, выпускаемых производственной системой, которая постоянно контролируется. Гистограмма данных очень похожа на нормальное распределение.

2. В табл. 16.4.1 приведены размеры прибыли фирм, занимающихся строительными материалами (по данным списка *Fortune 500*).

а) Постройте гистограмму значений прибыли (в процентах от продаж). Опишите вид распределения данных.

б) Найдите среднее значение и медиану. Объясните, почему они одинаковые (или разные).

в) Используя t-тест, установите, значимо ли отличается среднее значение прибыли (для идеализированной совокупности подобных фирм, работающих в таких же условиях) от значения опорной величины −5% (т.е. 5% потерь). (Используйте t-тест, даже если считаете, что он здесь не подходит.)

г) Используйте критерий знаков, чтобы решить вопрос о значимости отличия медианы прибыли этой идеализированной совокупности от 5% потерь.

д) Сравните эти два подхода к тестированию для указанных данных. В частности, объясните, какой из методов (критерий знаков или t-тест) более подходит в данном случае. Может, подходят оба метода? Почему?

3. В табл. 16.4.2 приведены размеры прибыли аэрокосмических фирм (по данным списка *Fortune 500*.)

Таблица 16.4.1. Прибыли фирм, занимающихся строительными материалами

Фирма	Прибыли (как процент от продаж)	Фирма	Прибыли (как процент от продаж)
American Standard	−1	Norton	7
Owens-Illinois	−2	Lafarge	7
Owens-Corning Fiberglas	7	Certainteed	4
USG	4	National Gypsum	−7
Manville	−59	Anchor Glass	−1
Corning Glass Works	10	Calmat	9
Nortek	1	Southdown	9

Данные взяты из *Fortune*, 1989, April 24, p. 380–381.

а) Постройте гистограмму размеров прибыли (в процентах от продаж). Опишите вид распределения.

б) Найдите среднее значение и медиану. Объясните, почему они одинаковые (или разные).

в) Используя t-тест, установите, значимо ли отличается среднее значение прибыли (для идеализированной совокупности подобных фирм, работающих в таких же условиях) от нуля. (Используйте t-тест, даже если считаете, что он здесь не подходит.)

г) Используйте критерий знаков, чтобы решить вопрос о значимости отличия медианы прибыли в этой идеализированной совокупности от нуля.

д) Сравните эти два подхода к тестированию для указанных данных. В частности, объясните, какой из методов (критерий знаков или t-тест) более подходит в данном случае. Может, подходят оба метода? Почему?

4. Среди 35 сотрудников вашего отдела продаж более половины имеют производительность выше медианы производительности по стране. Более точно, у 23 человек производительность выше, а у 12 — ниже. Можно ли утверждать, что это просто счастливая случайность, или производительность вашего отдела продаж действительно значимо выше медианы по стране? Как вы это определили?

5. Медиана количества телефонных звонков, которые ежедневно обрабатывал ваш отдел в прошлом году, равна 63 821. (Это медиана количества телефонных звонков, обрабатываемых ежедневно на протяжении года.) Пока в этом году больше половины дней количество звонков превышало этот уровень (было 15 дней с большим количеством звонков и 9 дней с меньшим количеством звонков). Можете ли вы утверждать, что по сравнению с прошлым годом вы перегружены телефонными звонками? Объясните, почему да или почему нет.

Таблица 16.4.2. Прибыли аэрокосмических фирм

Фирма	Прибыль (как процент от продаж)	Фирма	Прибыль (как процент от продаж)
United Technologies	4	Martin Marietta	6
Boeing	4	Grumman	2
McDonnell Douglas	2	Gencorp	3
Rockwell International	7	Sequa	4
Allied–Signal	4	Colt Industries	5
Lockheed	6	Sundstrand	−5
General Dynamics	4	Rohr Industries	4
Textron	3	Kaman	3
Northrop	2		

Данные взяты из *Fortune*, 1989, April 24, p. 380.

6. Реклама проходит тестирование на эффективность создания настроения расслабления. Выборка из 15 человек была опрошена до и после просмотра рекламного ролика. Вопросник включал много пунктов, но интересующий нас сейчас вопрос был сформулирован следующим образом: "Пожалуйста, опишите свое состояние в настоящий момент, используя шкалу от 1 (очень напряженное) до 5 (полностью расслабленное)". Результаты опроса приведены в табл. 16.4.3.

   а) Сколько человек ответило, что после просмотра рекламного ролика они стали чувствовать себя более расслабленными, чем до просмотра? Сколько человек указало на снижение расслабленности? У скольких настроение не изменилось?

   б) Найдите размер модифицированной выборки.

   в) Используйте непараметрический критерий знаков для разностей.

   г) Кратко обобщите полученные результаты с точки зрения эффекта этой рекламы (если он есть).

7. Исходя из идеи, что стресс говорящего неправду человека может быть измерен с помощью детектора лжи, был зафиксирован уровень стресса во время правдивых и ложных ответов шести испытуемых. Результаты приведены в табл. 16.4.4.

   а) Был ли у всех испытуемых уровень стресса выше при ложном ответе по сравнению с ответом правдивым?

Таблица 16.4.3. Влияние рекламы на настроение

Испытуемый	Оценка расслабленности	
	до	после
1	3	2
2	2	2
.	2	2
.	4	5
.	2	4
.	2	1
.	1	1
.	3	5
.	3	4
.	2	4
.	5	5
.	2	3
.	4	5
.	3	5
15	4	4

   б) У скольких людей уровень стресса был выше при правдивом ответе? У скольких при ложном ответе?

   в) Определите размер модифицированной выборки.

   г) Используя непараметрический критерий знаков для разностей, определите, значимо ли различаются уровни стресса при правдивом и ложном ответах.

8. Ваш отдел кадров направил 26 работников на консультацию по поводу злоупотребления алкоголем. Выяснилось, что 15 из них стали работать лучше, 4 — хуже, а остальные 7 человек продолжают работать на прежнем уровне. Используйте критерий знаков для разностей, чтобы дать ответ на вопрос, значимо ли больше людей улучшили свою работу по сравнению с теми, которые стали работать хуже.

9. Используйте данные задач 2 и 3 относительно объемов прибыли (как процента от продаж) аэрокосмических фирм и фирм, выпускающих строительные материалы.

   а) Определите медиану прибыли для каждой промышленной группы и сравните их.

   б) Объедините два набора данных, поместив в одной колонке названия промышленных групп, а в другой — процент прибыли фирмы.

   в) Сохраняя информацию о промышленной группе каждой фирмы, расположите значения прибыли (в процентах) в порядке возрастания. Добавьте третью колонку, в которую запишите общие ранги, усредняя соответствующим образом ранги одинаковых значений.

   г) Составьте список общих рангов по каждой промышленной группе.

   д) Определите средний ранг для каждой промышленной группы; вычислите разность между этими средними рангами (вычитая меньшее значение из большего).

   е) Определите соответствующую стандартную ошибку для этой разности средних рангов.

   ж) Вычислите тест-статистику непараметрического теста для двух независимых выборок.

   з) Из полученного результата проверки сделайте вывод относительно прибыли в этих двух промышленных группах.

Таблица 16.4.4. Уровень стресса

Испытуемый	Уровень стресса	
	правдивый ответ	ложный ответ
1	12,8	13,1
2	8,5	9,6
3	3,4	4,8
4	5,0	4,6
5	10,1	11,0
6	11,2	12,1

10. Вашей фирме предъявлен иск о дискриминации сотрудников по признаку пола, и вы изучаете документы, представленные другой стороной. Представленные ими данные приведены в табл. 16.4.5.

    а) Используя одинаковый масштаб, постройте блочные диаграммы для этих данных и прокомментируйте их.

    б) Используя непараметрический метод, проверьте, значимо ли различаются эти два распределения заработной платы.

    в) Исходя из результатов проверки сделайте краткое заключение.

11. Определяя конкурентоспособность вашей продукции, вы изучили ее надежность в сравнении с аналогичной продукцией ваших ближайших конкурентов. Каждое изделие подвергали жесткому испытанию таким образом, что износ изделия за день был примерно эквивалентен нормальной эксплуатации в течение года. В табл. 16.4.6 представлены данные испытаний на надежность.

Таблица 16.4.5. Данные о дискриминации сотрудников по признаку пола

Заработная плата, дол.	
женщины	мужчины
21 100	38 700
29 700	30 300
26 200	32 800
23 000	34 100
25 800	30 700
23 100	33 300
21 900	34 000
20 700	38 600
26 900	36 900
20 900	35 700
24 700	26 200
22 800	27 300
28 100	32 100
25 100	35 800
27 100	26 100
	38 100
	25 500
	34 000
	37 400
	35 700
	35 700
	29 100

Таблица 16.4.6. Надежность изделий при жесткой эксплуатации

Количество дней до поломки	
изделия вашей фирмы	изделия ваших конкурентов
1,0	0,2
8,9	2,8
1,2	1,7
10,3	7,2
4,9	2,2
1,8	2,5
3,1	2,6
3,6	2,0
2,1	0,5
2,9	2,3
8,6	1,9
5,3	1,2
	6,6
	0,5
	1,2

    а) Определите медиану времени до поломки для вашего изделия и для изделия ваших конкурентов. Вычислите разность медиан (вычитая меньшее значение из большего).

    б) Вычислите непараметрическую тест-статистику, чтобы определить, существенно ли отличается надежность вашего изделия от изделия конкурентов.

    в) Сформулируйте результат этой непараметрической проверки.

    г) Напишите краткое резюме, которое можно будет вставить в рекламную брошюру вашего изделия.

12. Есть ли препятствия в использовании непараметрического теста (две независимые выборки) для данных, приведенных в табл. 10.7.8 (см. главу 10), где указана месячная стоимость ухода за одним ребенком в детском саду в богатом районе Лорелхарст (Laurelhurst) в сравнении с другими районами Сиэтла? Почему?

13. Значимо ли различаются оценки качества вин "Шардоне" и "Каберне Совиньон", приведенные в табл. 10.7.6 (см. главу 10)? Имеем ли мы здесь дело со связанными или с независимыми выборками? Почему?

14. Число возвращенных изделий за каждый из последних 9 дней было следующим: 13, 8, 36, 18, 6, 4, 39, 47 и 21. Проверьте, значимо ли отличается медиана числа возвратов от 40, и определите p-значение (как одно из $p>0{,}05$, $p<0{,}05$ либо $p<0{,}01$).

15. Исходя из данных, приведенных в табл. 16.4.7, выполните непараметрический анализ цен лекарств в США и Канаде, выписанных по рецептам.

    а) Имеем мы здесь дело со связанными или с независимыми выборками?

    б) Значимо ли выше цены в Соединенных Штатах? Как вы это узнали?

Упражнения с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении А.

1. Используя непараметрический тест, выясните, значимо ли отличается медиана возраста служащих от 40 лет.
2. Используя непараметрический тест, выясните, значимо ли отличается медиана стажа работы служащих от трех лет.
3. Используя непараметрический тест, выясните, значимо ли отличается распределение размера годовой заработной платы мужчин от соответствующего распределения для женщин.

Проекты

1. Найдите в Internet, у своих друзей или в библиотеке данные об одном показателе, связанном с вашей работой или бизнес-интересами, и выберите приемлемое значение опорной величины для сравнения.

    а) Постройте для ваших данных гистограмму и прокомментируйте ее.

    б) Выполните t-тест и сделайте вывод.

    в) Используйте критерий знаков и сделайте вывод.

    г) Сравните результаты этих двух тестов. Если они различны, то результату какой из двух проверок следует доверять (если можно сделать подходящий выбор).

Таблица 16.4.7. Цены лекарств по рецептам (за 100 таблеток)

Лекарство	США	Канада
Ativan	49,43	6,16
Ceclor	134,18	84,14
Coumadin	36,70	19,59
Dilantin	15,03	4,67
Felden	167,54	123,61
Halcion	47,69	16,09
Lopressor	35,71	15,80
Naprosyn	72,36	42,64
Pepcid	103,74	76,22
Premarin	26,47	10,10

Данные взяты из *The Wall Street Journal*, 1993, February 16, p. A9. Источник. Prime Institute, University of Minnesota.

2. Продолжите работу с набором данных из предыдущего пункта, но введите в набор одно дополнительное значение, которое сильно отличается от всех остальных. Повторите пп. "а–г" предыдущего проекта для этого нового набора данных. Выполните дополнительно следующие задания.

    д) Опишите ваш опыт относительно того, как t-тест и критерий знаков реагируют на наличие такого экстремального значения.



3. Найдите в Internet, у своих друзей или в библиотеке два независимых набора данных об одном показателе, связанном с вашей работой или бизнес-интересами. Эти данные должны быть такими, чтобы имела смыл проверка значимости различий двух выборок.

    а) Используя один и тот же масштаб, постройте блочные диаграммы для ваших данных и прокомментируйте их.

    б) Запишите нулевую и исследовательскую (альтернативную) гипотезы для соответствующего непараметрического теста.

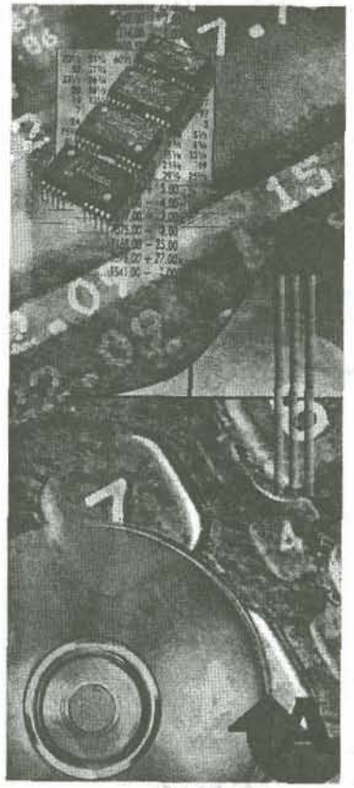
    в) Выполните соответствующий непараметрический тест и сделайте вывод.

    г) Используя результаты теста, напишите резюме об этой бизнес-ситуации.

ГЛАВА 17

Анализ "хи-квадрат": поиск закономерностей для качественных данных

Как можно сделать статистический вывод в отношении качественных данных, когда каждое наблюдение вместо числового значения представлено категорией (такой как цвет или источник энергии)? В двух случаях ответ нам известен. Во-первых, для *атрибутивных данных* (т.е. качественных данных, имеющих всего две категории) построение доверительных интервалов и проверка гипотез о процентах выполняются с использованием биномиального распределения и его нормальной аппроксимации. Во-вторых, для *порядковых данных* (когда значения имеют естественную упорядоченность) можно использовать непараметрические методы, изложенные в главе 16. Однако для *номинальных данных* (в которых отсутствует естественная упорядоченность) при наличии более чем двух категорий (или более одной переменной) необходимы другие методы. Ниже приведено несколько примеров.



Пример 1. Ни один производственный процесс, включая и ваш, не является совершенным. При наличии дефектов их группируют в категории в соответствии с вызвавшими их причинами. Общий процент дефектов определяется из *атрибутивной переменной* и может быть проанализирован с помощью биномиального распределения (в предположении о независимости). Процент бракованных изделий можно вычислить для каждой из причин, вызвавших дефект. Например, вы можете определить процент брака вследствие плохих микросхем, плохой пайки, некачественной монтажной платы и др. Поскольку каждую неделю каждый из этих процентов изменяется, хотелось бы знать, когда система выходит из-под контроля, отклоняясь на величину, большую, чем это может быть обусловлено только случайностью.

Пример 2. Опросы являются полезным источником информации. В дополнение к *подробностям политической жизни*, которым посвящены опросы средств массовой информации, многие фирмы также используют опросы, чтобы узнать, что их потребители (реальные и потенциальные) думают об истинном положении вещей и о перспективе. Эта информация полезна при планировании стратегии для маркетинга и внедрении нового изделия в производство. Многие опросы дают качественные данные, такие как категории "да", "нет", "нет ответа". Качественные данные также могут быть получены в результате выбора предпочтительного продукта из списка известных марок. В таком случае статистический вывод можно использовать для сравнения мнений двух групп людей, чтобы узнать, значимо ли различаются их взгляды. Или можно сравнить мнение некоторой одной группы с известным стандартом.

Критерий "хи-квадрат" используют для проверки гипотез о качественных данных, представленных не числами, а категориями. Для номинальных качественных данных можно только подсчитывать частоты (поскольку ранжирование или арифметические действия выполнять нельзя). Критерий (тест) "хи-квадрат" основан на частотах, которые представляют собой количество объектов выборки, попадающих в каждую из категорий. **Статистика "хи-квадрат"** измеряет разницу между *фактическими* частотами и *ожидаемыми* частотами (в предположении о справедливости нулевой гипотезы) следующим образом.

Статистика "хи-квадрат"

$$\text{Хи-квадрат статистика} = \text{Сумма}\ \frac{(\text{Наблюдаемая частота} - \text{Ожидаемая частота})^2}{\text{Ожидаемая частота}} = \sum \frac{(O_i - E_i)^2}{E_i},$$

где сумма вычисляется по всем категориям или комбинациям категорий. Определение ожидаемой частоты непосредственно зависит от того, какая именно гипотеза проверяется.

Используя статистику "хи-квадрат" в качестве меры того, насколько данные соответствуют нулевой гипотезе, критерий "хи-квадрат" позволяет принять решение о допустимости нулевой гипотезы.

17.1. Обобщение качественных данных с помощью частот и процентов

Ниже представлен типичный набор данных в виде списка результатов измерения для каждой из элементарных единиц выборки. В качестве элементарных единиц выступают люди, пришедшие в автосалон, а результатом измерения является тип предпочитаемого ими транспортного средства:

пикап, малолитражный автомобиль, малолитражный автомобиль, семейный седан, пикап, малолитражный автомобиль, спортивная машина, малолитражный автомобиль, семейный седан, пикап, малолитражный

автомобиль, автомобиль-фургон, автомобиль-фургон, малолитражный автомобиль, семейный седан, пикап, спортивная машина, семейный седан, семейный седан, малолитражный автомобиль, автомобиль-фургон, малолитражный автомобиль, семейный седан, спортивная машина, малолитражный автомобиль, малолитражный автомобиль, автомобиль-фургон, автомобиль-фургон...

Поскольку такой перечень может быть очень длинным, то, очевидно, что лучше работать с обобщающей таблицей частот или процентов. Таким образом, мы сохраняем всю содержащуюся в данных исходную информацию и в то же время представляем ее в более удобной и компактной форме. Примером может служить табл. 17.1.1.

Таблица 17.1.1. Предпочитаемая марка транспортного средства

Тип	Результат подсчета (частота)	Процент от общего количества
Семейный седан	187	(187/536) = 34,9
Малолитражный автомобиль	206	38,4
Спортивная машина	29	5,4
Автомобиль-фургон	72	13,4
Пикап	42	7,8
Общее количество	536	100,0

Обобщающая таблица частот или процентов также полезна при анализе *двумерных* качественных данных при наличии более одного измерения. Изучая отношение американцев к компании General Motors, *Business Week* работала с занимающейся опросами фирмой Louis Harris & Associates. Каждому респонденту было задано несколько вопросов[1]. В частности, рассматривались две следующие качественные переменные.

1. Ответ на вопрос: "Согласны ли вы с утверждением некоторых людей, что все автомобили General Motors выглядят одинаково?" Ответы классифицировались как "Согласен", "Не согласен", "Не уверен".
2. Отнесение респондентов на основании возраста и образования к одной из двух групп: "бэби-бумер" (baby boomer) (т.е. те, кто родились в период резкого увеличения рождаемости после Второй мировой войны) и другие (т.е. "не бэби-бумер"). Группа "бэби-бумер" определялась как "люди в возрасте от 18 до 39 лет с образованием не ниже колледжа".

Поскольку каждый респондент характеризуется категориями этих двух переменных, то фактический результат опроса можно представить в следующем виде:

несогласный "бэби-бумер", согласный "не бэби-бумер", несогласный "бэби-бумер", неуверенный "не бэби-бумер" и т.д.

Результаты опроса 1250 взрослых респондентов приведены в табл. 17.1.2.

Такая таблица частот или процентов помогает понять природу соответствующих качественных данных. Следующий шаг состоит в проверке различных гипотез в отношении этих частот и процентов.

[1] BW/Harris Poll: Americans Still Sold on General Motors", *Business Week*, 1987, March 16, p. 108.

Таблица 17.1.2. Ответы на вопрос об автомобилях компании General Motors

	Группа "бэби-бумер"	Группа "не бэби-бумер"	Итого
Согласен, %	52	39	42
Не согласен, %	43	53	51
Не уверен, %	5	8	7
Всего, %	100	100	100

Некоторые числа в этой таблице были рассчитаны исходя из оценки 20,3% для процента "бэби-бумер" во взрослом населении страны.

17.2. Проверка того, что значения процентов в генеральной совокупности равны некоторым заданным значениям

Нам уже известно, как с помощью биномиального распределения проверить равенство одного процента некоторой заданной опорной величины (см. главу 10). Однако для сравнения всей таблицы процентов с некоторой другой таблицей заданных опорных величин необходим другой метод. Достаточно распространенным применением такого рода теста является выяснение вопроса о том, является ли ваш нынешний опыт (выраженный в частотах и процентах) типичным по отношению к прошлому опыту (набор опорных величин).

Критерий "хи-квадрат" в отношении равенства процентов

Критерий "хи-квадрат" в отношении равенства процентов используют для проверки гипотезы о том, можно ли считать таблицу *наблюдаемых* частот или процентов извлеченной из некоторой генеральной совокупности с известным распределением процентов (известные опорные величины). Ниже сформулирована обобщенная задача и ее решение.

Критерий "хи-квадрат" в отношении равенства процентов

Данные: таблица частот для каждой категории одной качественной переменной.

Гипотезы:

H_0 : проценты в генеральной совокупности равны набору известных, фиксированных опорных величин.

H_1 : проценты в генеральной совокупности не равны этому набору опорных величин; есть отличие по крайней мере для одной категории.

Ожидаемые частоты: для каждой категории умножить известное значение доли в генеральной совокупности на размер выборки, n .

Предположения.

1. Набор данных представляет собой случайную выборку из рассматриваемой генеральной совокупности.
2. По крайней мере пять объектов ожидается в каждой категории.

Статистика "хи-квадрат":

$$\text{Хи-квадрат статистика} = \text{Сумма} \frac{(\text{Наблюдаемая частота} - \text{Ожидаемая частота})^2}{\text{Ожидаемая частота}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

Степени свободы: количество категорий минус единица.

Результат теста "хи-квадрат": статистически значим, если значение статистики "хи-квадрат" больше значения из табл. 17.2.1; в противном случае незначим.

Если значение статистики "хи-квадрат" *больше* критического значения из таблицы "хи-квадрат" для соответствующего числа степеней свободы, то это является свидетельством того, что наблюдаемые частоты значимо отличаются от тех, которые ожидаются исходя из известных вам опорных значений процентов. В этом случае следует отклонить нулевую гипотезу и принять исследовательскую (альтернативную) гипотезу, сделав вывод о том, что наблюдаемые выборочные проценты *значимо отличаются* от заданных опорных значений.

Таблица 17.2.1. Критические значения для теста "хи-квадрат"

Число степеней свободы	Уровень значимости 10%	Уровень значимости 5%	Уровень значимости 1%	Уровень значимости 0,1%
1	2,706	3,841	6,635	10,828
2	4,605	5,991	9,210	13,816
3	6,251	7,815	11,345	16,266
4	7,779	9.488	13,277	18,467
5	9,236	11,071	15,086	20,515
6	10,645	12,592	16,812	22,458
7	12,017	14,067	18,475	24,322
8	13,362	15,507	20,090	26,124
9	14,684	16,919	21,666	27,877
10	15,987	18,307	23,209	29,588
11	17,275	19,675	24,725	31,264
12	18,549	21,026	26,217	32,909
13	19,812	22,362	27,688	34,528
14	21,064	23,685	29,141	36,123
15	22,307	24,996	30,578	37,697
16	23,542	26,296	32,000	39,252
17	24,769	27,587	33,409	40,790
18	25,989	28,869	34,805	42,312

Число степеней свободы	Уровень значимости 10%	Уровень значимости 5%	Уровень значимости 1%	Уровень значимости 0,1%
19	27,204	30,144	36,191	43,820
20	28,412	31,410	37,566	45,305
21	29,615	32,671	38,932	46,797
22	30,813	33,924	40,289	48,268
23	32,007	35,172	41,638	49,728
24	33,196	36,415	42,980	51,179
25	34,382	37,652	44,314	52,620
26	35,563	38,885	45,642	54,052
27	36,741	40,113	46,963	55,476
28	37,916	41,337	48,288	56,892
29	39,087	42,577	49,588	58,301
30	40,256	43,733	50,892	59,703
31	41,422	44,985	52,191	61,098
32	42,585	46,194	53,486	62,487
33	43,745	47,400	54,776	63,870
34	44,903	48,602	56,061	65,247
35	46,059	49,802	57,342	66,619
36	47,212	50,998	58,619	67,985
37	48,363	52,192	59,893	69,346
38	49,513	53,384	61,162	70,703
39	50,660	54,572	62,428	72,055
40	51,805	55,758	63,691	73,402
41	52,949	56,942	64,950	74,745
42	54,090	58,124	66,206	76,084
43	55,230	59,304	67,459	77,419
44	56,369	60,481	68,710	78,749
45	57,505	61,656	69,957	80,077
46	58,641	62,830	71,201	81,400
47	59,774	64,001	72,443	82,720
48	60,907	65,171	73,683	84,037
49	62,038	66,339	74,919	85,351

Продолжение табл. 17.2.1

Число степеней свободы	Уровень значимости 10%	Уровень значимости 5%	Уровень значимости 1%	Уровень значимости 0,1%
50	63,167	67,505	76,154	86,661
51	64,295	68,669	77,386	87,968
52	65,422	69.832	78,616	89,272
53	66,548	70,993	79,843	90,573
54	67,673	72,153	81,069	91,872
55	68,796	73,311	82,292	93,167
56	69,919	74,468	83,513	94,461
57	71,040	75,624	84,733	95,751
58	72,160	76,778	85,950	97,039
59	73,279	77,931	87,166	98,324
60	74,397	79,082	88,379	99,607
61	75,514	80,232	89,591	100,888
62	76,630	81,381	90,802	102,166
63	77,745	82,529	92,010	103,442
64	78,860	83,675	93,217	104,716
65	79,973	84,821	94,422	105,988
66	81,085	85,965	95,626	107,258
67	82,197	87,108	96,828	108,526
68	83,308	88,250	98,028	109,791
69	84,418	89,391	99,228	111,055
70	85,527	90,531	100,425	112,317
71	86,635	91,670	101,621	113,577
72	87,743	92,808	102,816	114,835
73	88,850	93,945	104,010	116,091
74	89,956	95,081	105,202	117,346
75	91,061	96,217	106,393	118,599
76	92,166	97,351	107,583	119,850
77	93,270	98,484	108,771	121,100
78	94,374	99,617	109,958	122,348
79	95,476	100,749	111,144	123,594
80	96,578	101,879	112,329	124,839

Окончание табл. 17.2.1

Число степеней свободы	Уровень значимости 10%	Уровень значимости 5%	Уровень значимости 1%	Уровень значимости 0,1%
81	97,680	103,010	113,512	126,083
82	98,780	104,139	114,695	127,324
83	99,880	105,267	115,876	127,565
84	100,980	106,395	117,057	129,804
85	102,079	107,522	118,236	131,041
86	103,177	108,648	119,414	132,277
87	104,275	109,773	120,591	133,512
88	105,372	110,898	121.767	134,745
89	106,469	112,022	122,942	135,978
90	107,565	113,145	124,116	137,208
91	108,661	114,268	125,289	138,438
92	109,756	115,390	126,462	139,666
93	110,850	116,511	127,633	140,893
94	111,944	117,632	128,803	142,119
95	113,038	118,752	129,973	143,344
96	114,131	119,871	131,141	144,567
97	115,223	120,990	132,309	145,789
98	116,315	122,108	133,476	147,010
99	117,407	123,225	134,642	148,230
100	118,498	124,342	135,807	149,449

Если значение статистики "хи-квадрат" *меньше* критического значения из таблицы "хи-квадрат", то наблюдаемые значения не очень отличаются от значений, которые можно ожидать исходя из известных опорных значений процентов. В этом случае следует принять нулевую гипотезу (как приемлемую возможность) и сделать вывод, что наблюдаемые выборочные проценты *не имеют значимых отличий* от заданных опорных значений.

Грубое эмпирическое правило гласит, что ожидаемые частоты в каждой категории должны быть по крайней мере не меньше пяти, поскольку тест "хи-квадрат" является приблизительным, а не точным тестом. Если сформулированное в этом правиле требование выполняется, то этой аппроксимации вполне достаточно для практических целей, но можно получить ошибочный результат, если ожидаемые частоты для некоторых категорий слишком малы. Риск состоит в том, что в таком случае нельзя контролировать на уровне 5% (или на любом другом выбранном уровне) вероятность ошибки первого рода.

Пример. *Три причины наличия проблем с качеством*

Частью обязательств вашей фирмы в отношении тотального контроля качества является тщательная регистрация всех дефектов, так как это позволяет получить полезную информацию для улучшения качества. Каждый некачественный компонент проверяют, чтобы установить причину дефекта — плохая микросхема, плохая пайка или плохая монтажная плата. Исходя из данных о работе этой сборочной линии в прошлом, вам известны ожидаемые проценты брака (опорные проценты) для случая, когда производственный процесс находится под контролем. Сравнивая текущие значения с этими опорными процентами, можно проверить, находится сейчас процесс под контролем или нет[2].

В табл. 17.2.2 содержатся данные о причинах брака за прошедшую неделю.

В табл. 17.3.2 приведены значения опорных величин, взятые из данных прошлых лет, когда сборочная линия работала надлежащим образом.

Хотя количество некачественных микросхем на прошлой неделе (16%) было близко к опорному значению (15,2%), другие показатели отличаются от соответствующих опорных величин достаточно сильно (например, 70% по сравнению с 60,5% для некачественной пайки). Вопрос заключается в том, значима ли эта разница? Другими словами, могут ли полученные на прошлой неделе значения объемов брака рассматриваться как результат извлечения случайной выборки из генеральной совокупности, в которой проценты брака соответствуют опорным величинам? Или вопрос можно сформулировать так, достаточно ли велика наблюдаемая разница, чтобы ее нельзя было объяснить только случайностью? Тест "хи-квадрат" равенства процентов даст ответ на этот вопрос. Гипотезы формулируются следующим образом.

- H_0 : процесс все еще под контролем. (Наблюдаемые объемы брака равны опорным величинам.)
- H_1 : процесс вышел из-под контроля. (Наблюдаемые объемы брака не равны опорным величинам.)

Таблица 17.2.2. Наблюдаемые данные о бракованных компонентах за прошедшую неделю

Проблема	Наблюдаемое значение (частота)	Процент от общего количества
Микросхема	8	(8/50=) 16
Пайка	35	70
Плата	7	14
Итого	50	100

Таблица 17.2.3. Опорные значения процентов бракованных компонентов из данных прошлых лет, когда процесс находился под контролем

Проблема	Процент от общего количества
Микросхема	15,2
Пайка	60,5
Плата	24,3
Итого	100

[2] Конечно, особо следует также обратить внимание на общий процент брака. Представленный здесь анализ помогает выявить только проблемы определенного вида. В целом тема контроля качества будет рассмотрена в главе 18.

Таблица 17.2.4. Ожидаемые частоты: предполагаемое количество бракованных компонентов, вычисленное в соответствии с опорными значениями процентов для процесса, находящегося под контролем

Проблема	Ожидаемое значение
Микросхема	(0,152 × 50=) 7,60
Пайка	30,25
Плата	12,15
Итого	50

В табл. 17.2.4 содержатся ожидаемые частоты, вычисленные путем умножения значений опорных величин процентов (15,2%, 60,5% и 24,3%) на размер выборки $n = 50$. Заметим, что это вполне нормально, что вычисленные значения ожидаемых частот содержат дробную часть; это необходимо для того, чтобы вычисленные значения действительно точно соответствовали опорным значениям. Обратите также внимание, что сумма всех ожидаемых частот равна сумме всех реальных (наблюдаемых) частот, а именно размеру выборки $n = 50$.

Что касается выполнения необходимых допущений, то этот набор данных взят из идеализированной генеральной совокупности всех компонентов, которые можно было бы произвести в аналогичных условиях. Таким образом, имеющиеся данные рассматривают как случайную выборку из идеализированной генеральной совокупности возможных результатов. Второе допущение также выполняется, поскольку все ожидаемые частоты по крайней мере не меньше 5 (все значения 7,6; 30,3 и 12,2 больше 5). Обратите внимание, что это допущение проверяют в отношении *ожидаемых*, а не *наблюдаемых* частот.

Статистику "хи-квадрат" вычисляют исходя из наблюдаемых и ожидаемых частот для всех категорий следующим образом (конечно, общая сумма не рассматривается, поскольку это не категория):

$$\text{Хи-квадрат статистика} = \sum \frac{(\text{Наблюдаемая частота} - \text{Ожидаемая частота})^2}{\text{Ожидаемая частота}} =$$

$$= \frac{(8-7{,}60)^2}{7{,}60} + \frac{(35-30{,}25)^2}{30{,}25} + \frac{(7-12{,}15)^2}{12{,}15} =$$

$$= \frac{0{,}1600}{7{,}60} + \frac{22{,}5625}{30{,}25} + \frac{26{,}5225}{12{,}15} = 0{,}0211 + 0{,}7459 + 2{,}1829 = 2{,}950.$$

Число степеней свободы на единицу меньше, чем число категорий. Здесь у нас три категории (микросхема, пайка, плата), следовательно,

число степеней свободы равно: 3 – 1 = 2.

Используя таблицу "хи-квадрат", определяем, что для двух степеней свободы и проверки на уровне 5% критическое значение равно 5,59. Поскольку значение статистики "хи-квадрат" (2,950) меньше табличного значения, то нулевая гипотеза принимается. Таким образом, пропорция объемов брака различного типа на сборочной линии соответствует опорным значениям для линии, которая находится под контролем, а имеющиеся расхождения обусловлены только лишь случайностью для выборки размером 50. У вас нет убедительных доказательств того, что процесс вышел из-под контроля, поэтому вы соглашаетесь с тем, что процесс продолжает находиться под контролем.

Наблюдаемые значения процентов незначимо отличаются от опорных значений. Исходя из этого у нас нет оснований считать, что процесс вышел из-под контроля.

17.3. Проверка взаимосвязи между двумя качественными переменными

Предположим, что имеются две качественные переменные, т.е. набор данных представляет собой *двумерные качественные данные*. После изучения каждой переменной отдельно с помощью анализа частот и процентов вас может заинтересовать вопрос о *связи* (если она есть) между этими переменными. В частности, связаны ли вообще каким-либо образом между собой две переменные. Ниже приведены ситуации, когда такой тест может быть полезен.

1. Одна переменная представляет собой предпочитаемый человеком вид досуга (выбранный из перечня, включающего спорт, просмотр телевизора, чтение и т.д.). Другая переменная представляет собой любимый вид сухого завтрака человека. Чтобы лучше спланировать маркетинговую стратегию, желательно понять связь между этими переменными. Если вы работаете в сфере производства сухих завтраков, то такой анализ поможет решить, что именно положить в пакет для завтрака. Если вы работаете в сфере организации досуга, то анализ поможет решить, какие компании, производящие сухие завтраки, можно привлечь к выработке совместных маркетинговых планов.
2. Одна переменная фиксирует причину брака в некотором изделии. Другая переменная представляет собой имя менеджера, ответственного за производство этого изделия. Знание связи (если она есть) между этими двумя переменными поможет сконцентрировать усилия на том, чтобы определить конкретного менеджера, которому следует уделять больше внимания решению проблем качества. Если эта причина брака характерна для всех менеджеров, то это систематическая проблема, и к ее решению следует подходить более широко (не с точки зрения отдельного менеджера). Если же эта причина брака характерна только для одного менеджера, то следует заставить именно этого менеджера заняться решением выявленной проблемы.

Понятие независимости переменных

Говорят, что две качественные переменные являются **независимыми**, если знание значения одной переменной не помогает предсказать значение другой. Другими словами, *вероятности* для одной переменной такие же, как и *условные вероятности* при заданных значениях другой переменной. Каждая переменная характеризуется собственными значениями процентов, которые представляют собой вероятности появления каждой из категорий. **Условные проценты в генеральной совокупности** представляют собой вероятности появления категорий одной переменной при ограничении рассмотрения только одной категории другой переменной. Такие ограниченные проценты в генеральной совокупности представляют собой условные вероятности одной переменной для этой заданной категории другой переменной.

Например, представим, что в генеральной совокупности процент брака "отслаивание краски" равен 3,1%. Однако, когда работает менеджер Джонс, условный процент брака "отслаивание краски" в генеральной совокупности равен 11,2%. В этом случае знание значения одной переменной (имя конкретного менеджера) помогает спрогнозировать значение другой переменной (объем брака определенного типа), поскольку 3,1% и 11,2% различаются между собой. Появление брака "отслаивание краски" более вероятно во время работы Джонса и менее вероятно, когда работает кто-то другой. Следовательно, эти две переменные *не являются независимыми*.

Обратите внимание, что реальная ситуация обычно сложнее этого примера, поскольку приходится работать с *выборочными процентами*, которые являются оценками вероятностей в генеральной совокупности. Нельзя просто анализировать эти проценты и смотреть, различаются ли они, потому что благодаря случайности они всегда (или почти всегда) различаются. Тест "хи-квадрат" на независимость переменных позволяет установить, когда различия *выходят за рамки* тех, которые можно было бы ожидать исходя только из случайности.

Критерий "хи-квадрат" независимости

Критерий "хи-квадрат" независимости используют для решения вопроса о наличии связи между двумя качественными переменными исходя из таблицы наблюдаемых частот для двумерного набора качественных данных. Критерий использует таблицу частот, которые можно было бы ожидать в случае, если переменные независимы. Ниже приведена формулировка задачи и ее решение.

Критерий "хи-квадрат" независимости

Данные: таблица частот всех комбинаций категорий двух качественных переменных, созданная для некоторого двумерного набора данных.

Гипотезы.

- H_0 : две переменные не зависят одна от другой. Другими словами, вероятности распределения каждой из переменной равны условным вероятностям, которые определяются другой переменной.
- H_1 : две переменные связаны; они не являются независимыми друг от друга. Существует по крайней мере одна категория одной переменной, чья вероятность не равна условной вероятности для некоторой определенной категории другой переменной.

Таблица ожидаемых частот: для каждой комбинации двух категорий (категорий разных переменных) частоту одной категории умножаем на частоту другой категории и полученное произведение делим на общий размер выборки n :

$$\text{Ожидаемая частота} = \frac{\left(\begin{array}{c}\text{Частота категории}\\ \text{для одной переменной}\end{array}\right)\left(\begin{array}{c}\text{Частота категории}\\ \text{для другой переменной}\end{array}\right)}{n}$$

Допущения.

1. Набор данных представляет собой случайную выборку из рассматриваемой генеральной совокупности.
2. Для каждой комбинации категорий ожидаемая частота по крайней мере не меньше пяти.

Статистика "хи-квадрат":

$$\text{Хи-квадрат статистика} = \sum \frac{(\text{Наблюдаемая частота} - \text{Ожидаемая частота})^2}{\text{Ожидаемая частота}},$$

где суммирование производится по всем комбинациям категорий.

Степени свободы:

(Число категорий для первой переменной – 1)(Число категорий для второй переменной – 1).

Результат теста "хи-квадрат": наличие значимой связи, если значение статистики "хи-квадрат" больше табличного значения; в противном случае значимой связи нет.

Если значение статистики "хи-квадрат" больше критического значения из таблицы "хи-квадрат" (табл. 17.2.1), то получено доказательство, что наблюдаемое значение частот намного отличается от тех, которых можно было бы ожидать в случае, если переменные были бы независимыми. В такой ситуации следует отклонить нулевую гипотезу о независимости и принять исследуемую (альтернативную) гипотезу. Можно сделать вывод о том, что между переменными существует *значимая связь*, или, другими словами, они *не являются независимыми* друг от друга.

Если значение "хи-квадрат" статистики меньше критического значения из таблицы "хи-квадрат", то наблюдаемые данные ненамного отличаются от тех, которых можно было бы ожидать, если переменные в генеральной совокупности были бы независимыми одна от другой. Следует принять нулевую гипотезу о независимости переменных как приемлемую возможность. Можно сделать вывод, что между переменными отсутствует значимая связь. Это слабое заключение, поскольку принята нулевая гипотеза о независимости переменных. Вы *признали* независимость переменных, но вы ее *не доказали*.

Почему именно так вычисляют ожидаемые частоты? Вспомним, что в соответствии с теорией вероятности (см. главу 6) для двух независимых событий вероятность того, что *оба* они наступят, равна произведению вероятностей этих событий. Уравнение, которое определяет ожидаемую частоту, выражает независимость, по сути, через умножение этих вероятностей[3].

Пример. Сегментирован ли ваш рынок

Вы пытаетесь разработать стратегию для проведения маркетинговой кампании новой линии продукции, состоящей из трех спортивных тренажеров, имитирующих греблю на лодке. Базовая модель сделана из прочного хрома с черными пластиковыми деталями и квадратным сиденьем. Модель с улучшенным дизайном представлена разнообразием цветов и рельефным сиденьем. Полная модель получена за счет добавления ряда аксессуаров к улучшенной модели (компьютеризованный дисплей, звук падающей воды, другие звуковые эффекты и т.п.).

[3] Чтобы убедиться, что это действительно так, достаточно разделить обе части уравнения на n. В результате получим

$$\frac{\text{Ожидаемая частота}}{n} = \frac{\begin{pmatrix}\text{Частота категории}\\ \text{одной переменной}\end{pmatrix}}{n} \times \frac{\begin{pmatrix}\text{Частота категории}\\ \text{другой переменной}\end{pmatrix}}{n}.$$

Поскольку деление на n дает долю, которая оценивает вероятность, то уравнение показывает, что вероятность комбинации конкретной категории одной переменной и конкретной категории другой переменной равна произведению вероятностей этих категорий. Это ничем не отличается от определения независимости для вероятностей (см. главу 6).

Чтобы помочь отделу написать информационную брошюру и пресс-релиз, вам необходимо узнать, какой тип модели предпочитает каждый из типов покупателей. Например, вам не нужно лезть из кожи вон[4], показывая практичность своей модели, в то время как ваш рынок фактически состоит из импульсивных клиентов.

Маркетинговая фирма собрала данные на небольшом пробном рынке. Для каждой покупки фиксировали две качественные переменные. Одна переменная представляла тип модели (базовая, улучшенная, полная), а другая — тип потребителя (потребителя характеризовали как практичного или импульсивного). В табл. 17.3.1 приведен набор данных, представляющий собой таблицу частот для этих $n = 221$ потребителей. Например, из 221 покупки 22 базовые модели приобретены практичными клиентами.

Общие проценты, полученные путем деления значения частот на размер выборки, n, показывают, какой процент соответствует каждой категории каждой из переменных и каждой комбинации таких категорий (в комбинации используется по одной категории из каждой переменной). Из данных табл. 17.3.2 видно, что наибольшую группу составляют продажи улучшенной модели импульсивным покупателям (39,8% от всех продаж).

Следующую по размеру группу представляют продажи полной модели практичным покупателям (24,4% от общего числа продаж). Из итоговых процентов для типа модели (крайняя правая колонка) видно, что наименее продаваемой является базовая модель (только 21,3% проданных тренажеров составляют базовые модели). Полученные результаты свидетельствуют о том, что у вас обеспеченные клиенты, готовые платить больше за модели высокого качества. (Это хорошие новости. Поздравляю!)

Проценты по моделям, полученные путем деления каждой из частот на общую сумму частот для этой модели, показывают процент покупателей каждого типа для каждой из моделей. Это *условные выборочные проценты для заданного типа модели*, которые оценивают соответствующие условные проценты генеральной совокупности (условные вероятности). Так, вычисленные проценты показывают профиль покупателя для каждого типа тренажеров. Из данных табл. 17.3.2 видно, что базовую модель покупают примерно в равных пропорциях оба типа покупателей (46,8% покупок сделано практичными покупателями и 53,2% — импульсивными). Улучшенную модель приобретают почти исключительно импульсивные потребители, в то время как покупатель полной модели, скорее, практичный, чем импульсивный.

Проценты по типам покупателей, полученные путем деления каждой из частот на сумму частот для данного типа покупателей, показывают проценты покупок модели каждого типа конкретным типом потребителя. Это *условные выборочные проценты для заданного типа потребителя*, которые оценивают соответствующие условные проценты генеральной совокупности (условные вероятности). Полученный результат показывает профиль предпочтений модели для каждого из типов потребителей. Из данных табл. 17.3.4 видно, что практичные потребители определенно предпочитают полную модель (60,7% покупателей такого типа приобрели эту модель), а импульсивные покупатели определенно предпочитают улучшенную модель (66,7% их приобрели эту модель). Однако нельзя игнорировать и другие варианты выбора (как, например, покупку базовой модели практичными покупателями), поскольку они также представляют значимый, хотя и меньший сегмент вашего рынка.

Похоже ли, что эти две переменные независимы? Нет. Мы уже отметили несколько фактов, указывающих на некоторую связь между типом клиента и типом предпочитаемой им модели. Например, из таблицы процентов по типам покупателей (табл. 17.3.4) видно, что полную модель приобрели 33% всех покупателей и намного больший процент (60,7%) практичных покупателей. Если бы эти две переменные были независимы, то следовало бы ожидать, что практичные покупатели будут демонстрировать те же предпочтения. Таким образом, похоже, что знание потребителя действительно помогает в прогнозировании того, какая модель будет им куплена, что, в свою очередь, предполагает, что эти два фактора не являются независимыми.

Если бы переменные были независимы, то проценты по типам клиентов были бы одинаковыми во всех трех колонках: практичные покупатели и импульсивные покупатели имели бы одинаковый профиль покупок моделей, который совпадал бы с профилем всех покупателей в целом. Аналогично, если бы переменные были независимы, то проценты по моделям были бы одинаковыми во всех четырех строках: базовая модель, улучшенная модель и полная модель имели бы одинаковые профили типов покупателей, совпадающий с профилем типа покупателей для всех моделей вместе.

[4] Простите за каламбур.

Таблица 17.3.1. Частоты: покупки тренажеров

	Практичный покупатель	Импульсивный покупатель	Итого
Базовая модель	22	25	47
Улучшенная	13	88	101
Полная	54	19	73
Итого	89	132	221

Таблица 17.3.2. Общие проценты: покупки тренажеров

	Практичный покупатель, %	Импульсивный покупатель, %	Итого, %
Базовая	(22/221=) 10,0	11,3	21,3
Улучшенная	5,9	39,8	45,7
Полная	24,4	8,6	33
Итого	40,3	59,7	100

Таблица 17.3.3. Проценты по моделям: покупки тренажеров

	Практичный покупатель, %	Импульсивный покупатель, %	Итого, %
Базовая	(22/47=) 46,8	53,2	100
Улучшенная	12,9	87,1	100
Полная	74,0	26,0	100
Итого	40,3	59,7	100

Таблица 17.3.4. Проценты по типам покупателей: покупки тренажеров

	Практичный покупатель, %	Импульсивный покупатель, %	Итого, %
Базовая	(22/89) = 24,7	18,9	21,3
Улучшенная	14,6	66,7	45,7
Полная	60,7	14,4	33
Итого	100	100	100

Таблица 17.3.5. Ожидаемые частоты: покупки тренажеров

	Практичный покупатель	Импульсивный покупатель	Итого
Базовая	(89×47/221 =) 18,93	28,07	47
Улучшенная	40,67	60,33	101
Полная	29,40	43,60	73
Итого	89	132	221

Ожидаемая таблица (полученная умножением итоговой суммы частот для каждого покупателей на итоговую сумму частот для каждой модели и последующим делением полученного произведения на общий размер выборки $n = 221$) показывает частоты покупок, которых следовало бы ожидать, если бы тип покупаемой модели был независим от типа покупателя. Из табл. 17.3.5 видно, что итоговая сумма ожидаемых частот остается той же, что и в наблюдаемой таблице (89 для практичных покупателей и 132 для импульсивных). То же самое для количества покупок каждой из моделей (47 базовых, 101 улучшенная и 73 полных). Но частоты внутри таблицы перераспределились, показывая, чего можно было бы ожидать (в среднем) в случае справедливости предположения о независимости переменных.

Обратите внимание, можно было бы ожидать, что практичные покупатели купят 40,67 тренажеров улучшенной модели (поскольку количество практичных покупателей составляет 89 человек и всего был продан 101 тренажер улучшенной модели из общего числа (221) проданных тренажеров). Однако фактически продано только 13 (см. исходную таблицу наблюдаемых частот), что намного меньше количества 40,67, ожидаемого исходя из предположения о независимости переменных.

Является ли тип покупаемой модели независимым от типа покупателя? Действительно ли разница между ожидаемыми и фактическим значениями больше, чем та, которая могла бы образоваться только благодаря случайности, если бы переменные действительно были независимыми? Ответ на этот вопрос даст критерий "хи-квадрат".

Соответствующие гипотезы формулируются следующим образом:

- H_0 : тип покупателя не зависит от покупаемой модели.
- H_1 : тип покупателя не является независимым от типа покупаемой модели.

Нулевая гипотеза утверждает, что все покупатели имеют одинаковые предпочтения (процент покупок каждой из моделей) независимо от того, какому типу они принадлежат (практичный покупатель или импульсивный). Исследуемая (альтернативная) гипотеза утверждает, что эти предпочтения различаются.

Давайте проверим предположения. Является ли этот набор данных случайной выборкой из интересующей нас генеральной совокупности? В действительности нет, но он может быть достаточно близким к такой выборке. Это частично зависит от того, насколько тщательно было спланировано это маркетинговое исследование, т.е. насколько представителен этот пробный рынок по отношению ко всем вашим покупателям. Помните, что статистический вывод может быть распространен только на такую большую генеральную совокупность (реальную или идеализированную), которая *представлена* вашей выборкой (иными словами, для которой ваша выборка является репрезентативной). Поэтому для нашего случая давайте будем считать, что набор данных представляет собой случайную выборку из генеральной совокупности покупок, сделанных в городах, аналогичных той зоне, в которой проводилось исследование, и в магазинах, подобных тем, которые были включены в исследование. Второе необходимое допущение выполняется, поскольку все значения в ожидаемой таблице не меньше 5.

Значение статистики "хи-квадрат" равно сумме термов (наблюдаемая — ожидаемая)^2/ожидаемая. Эти значения приведены в табл. 17.3.6 (обратите внимание, для итоговой строки и итоговой колонки вычисления не производятся):

$$\text{Хи-квадрат статистика} = \sum \frac{(\text{Наблюдаемая частота} - \text{Ожидаемая частота})^2}{\text{Ожидаемая частота}} =$$
$$= 0,50+18,83+20,58+0,34+12,69+13,88 = 66,8.$$

Число степеней свободы равно двум, поскольку у нас три категории моделей и две категории покупателей:

$$\text{Число степеней свободы} = (3-1)(2-1) = 2\times 1 = 2.$$

В таблице "хи-квадрат" для двух степеней свободы и уровня значимости 5% находим критическое значение, равное 5,991. Поскольку значение "хи-квадрат" статистики (66,8) больше этого критического значения, то между этими двумя качественными переменными имеется значимая связь. Ввиду того что значение статистики "хи-квадрат" так велико, давайте проверим гипотезу на уровне значимости 0,1%.

Таблица 17.3.6. (наблюдаемая — ожидаемая)^2/ожидаемая: покупки тренажеров

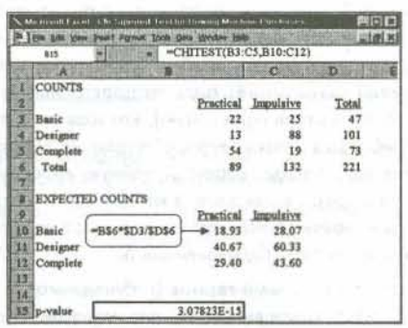
	Практичный покупатель	Импульсивный покупатель
Базовая	$\left([22-18{,}93]^2/8{,}93=\right)0{,}50$	0,34
Улучшенная	18,83	12,69
Полная	20,58	13,88

В этом случае критическое значение (для двух степеней свободы) равно 13,816. Значение статистики "хи-квадрат" и в этом случае значительно выше. На основании этого делаем следующий вывод.

Связь между типом покупателя и типом покупаемой модели является очень высоко значимой ($p<0{,}001$).

Поскольку в предположении о независимости вероятность того, что вы получите данные со столь высокой связью так мала ($p<0{,}001$), вы располагаете очень убедительным свидетельством против нулевой гипотезы о независимости переменных. Теперь вы можете планировать маркетинговую кампанию со значительной уверенностью в том, что различные модели тренажеров действительно имеют разную привлекательность в разных сегментах рынка.

Вычислить p -значение для теста "хи-квадрат" независимости можно с помощью функции `=CHITEST()` (`=ХИ2ТЕСТ()`) электронной таблицы Excel, но сначала необходимо вычислить таблицу ожидаемых частот. Ниже показаны результаты вычислений: сначала исходная таблица частот, затем таблица ожидаемых частот[5] и, наконец, функция `=CHITEST()` (`=ХИ2ТЕСТ()`), которая использует обе эти таблицы. В результате вычислений с помощью функции `=CHITEST()` (`=ХИ2ТЕСТ()`) получаем p -значение, равное `3,07823E-15`, т.е. очень маленькое число `0,000000000000003078 23`, поскольку запись `E-15` означает необходимость перенести десятичную точку на 15 позиций влево. Таким образом, результат является очень высоко значимым, так как p -значение меньше 0,001.



Начните с выбора ячейки, в которой вы хотите получить p -значение. Затем выберите функцию Insert⇨Function (Вставка⇨Функция) из главного меню, далее выберите в качестве категории функций Statistical (Статистические) и в качестве имени функции — `CHITEST` (`ХИ2ТЕСТ`). Откроется диалоговое окно, в котором сначала перетащите курсор, чтобы указать таблицу частот, затем щелкните на Expected range box (Ожидаемый интервал), перетащите курсора, чтобы указать и таблицу ожидаемых частот, нажмите <Enter> для завершения процесса вычислений. Вот как это выглядит.



[5] Чтобы создать формулу вычисления ожидаемых частот, которая затем будет корректно копироваться для заполнения всей таблицы, используйте знак доллара для задания "абсолютного адреса" в формуле `=B$6*$D3/$D$6` при вычислении ожидаемого количества (18,93) покупок базовых моделей тренажеров практичными покупателями. Эту формулу можно скопировать и вставлять в таблицу при ее заполнении, при этом всегда будут использоваться итоговые суммы из строки 6 (поэтому здесь указана ссылка `B$6`), итоговые суммы из колонки D (поэтому здесь указана ссылка `$D3`) и общая итоговая сумма из ячейки D6 (поэтому здесь указана ссылка `$D$6`).

17.4. Дополнительный материал

Резюме

Качественные данные анализируют в терминах частот и процентов. **Критерий "хи-квадрат"** используют для проверки гипотез о качественных данных, выраженных категориями, а не числовыми значениями. **Статистика "хи-квадрат"** измеряет разницу между *фактическими* частотами и *ожидаемыми* (при справедливости нулевой гипотезы) частотами.

$$\text{Хи-квадрат статистика} = \sum \frac{(\text{Наблюдаемая частота} - \text{Ожидаемая частота})^2}{\text{Ожидаемая частота}},$$

где сумма вычисляется по всем категориям или комбинациям категорий. Определение *ожидаемой частоты* зависит от формулировки конкретной нулевой гипотезы, которая тестируется.

Тест "хи-квадрат" в отношении равенства процентов используют для проверки гипотезы о том, что таблица *наблюдаемых* частот или процентов (характеризующая одну качественную переменную) построена на данных из некоторой генеральной совокупности с известными значениями процентов (опорными величинами). Гипотезы формулируются следующим образом.

- H_0 : значения процентов в генеральной совокупности равны набору известных, заданных опорных величин.
- H_1 : значения процентов в генеральной совокупности не равны заданному набору опорных величин. По меньшей мере для одной категории есть различие.

Допущения.

1. Набор данных представляет собой случайную выборку из изучаемой генеральной совокупности.
2. Ожидается наличие по крайней мере пяти объектов в каждой из категорий.

В статистике "хи-квадрат" ожидаемое значение частоты для каждой категории равно произведению заданного опорного значения процента в генеральной совокупности на размер выборки n . Число степеней свободы равно количеству категорий минус один.

Если значение "хи-квадрат" статистики больше критического значения из таблицы "хи-квадрат" для соответствующего числа степеней свободы, то это служит доказательством того, что наблюдаемые частоты сильно отличаются от частот,

ожидаемых исходя из опорных значений процентов. В таком случае следует отклонить нулевую гипотезу и принять исследуемую (альтернативную) гипотезу. Наблюдаемые выборочные проценты значимо отличаются от опорных значений.

Если значение "хи-квадрат" статистики меньше критического значения из таблицы "хи-квадрат", то наблюдаемые частоты не сильно отличаются от частот, которых можно было бы ожидать исходя из опорных значений процентов. В таком случае следует принять нулевую гипотезу как допустимую. Наблюдаемые выборочные проценты *не* имеют значимых отличий от опорных значений.

При наличии *двумерного набора качественных данных* можно проверить наличие связи между двумя переменными. Говорят, что две качественные переменные являются **независимыми**, если знание значения одной из них не помогает предсказать значение другой; другими словами, *вероятности* одной переменной такие же, как *условные вероятности* при заданных значениях другой переменной. **Условные проценты генеральной совокупности** представляют собой вероятности появления значений одной переменной при ограничении рассмотрения только одной категории другой переменной. Выборочный набор данных дает оценки процентов генеральной совокупности и условных процентов генеральной совокупности.

Один из способов анализа качественных данных состоит в использовании *общих процентов*, что позволяет определить относительную частоту каждой *комбинации* пар категорий (по одной для каждой переменной). Другой способ заключается в использовании *процентов по одной из переменных* с целью получить профиль оценок условных вероятностей другой переменной для каждой из категорий этой первой переменной.

Критерий "хи-квадрат" независимости используют для выяснения, являются ли две переменные независимыми или нет, исходя из таблицы наблюдаемых частот, созданной для двумерного набора качественных данных. Критерий вычисляют исходя из таблицы частот, которых следовало бы ожидать в случае независимости двух переменных. Гипотезы формулируются следующим образом:

- H_0: две переменные не зависят одна от другой. Другими словами, вероятности одной переменной равны условным вероятностям при фиксированных значениях другой переменной.
- H_1: две переменные связаны; другими словами, они не являются независимыми друг от друга. Существует по крайней мере одна категория одной переменной, вероятность которой не равна условной вероятности при фиксированной категории другой переменной.

Таблицу ожидаемых частот строят следующим образом: для каждой комбинации категорий (по одной категории от каждой из переменных) частоту одной категории умножают на частоту другой категории и полученное произведение делят на общий размер выборки n.

$$\text{Ожидаемая частота} = \frac{\left(\begin{array}{c}\text{Частота категории}\\ \text{для одной переменной}\end{array}\right)\left(\begin{array}{c}\text{Частота категории}\\ \text{для другой переменной}\end{array}\right)}{n}.$$

Допущения.

1. Набор данных представляет собой случайную выборку из изучаемой совокупности.
2. Для каждой комбинации категорий ожидаемая частота не меньше пяти.

При вычислении статистики "хи-квадрат" для проверки гипотезы о независимости число степеней свободы рассчитывают по следующей формуле.

$$\begin{aligned}&\text{Число степеней свободы} =\\ &= (\text{Число категорий для первой переменной} - 1)\times\\ &\times(\text{Число категорий для второй переменной} - 1)\end{aligned}$$

Если значение статистики "хи-квадрат" больше критического значения из таблицы "хи-квадрат", то получено доказательство того, что наблюдаемые значения частот сильно отличаются от тех, что имели бы место, если бы переменные были независимы. В таком случае следует отклонить нулевую гипотезу о независимости переменных и принять исследуемую (альтернативную) гипотезу, сделав вывод, что между переменными имеется значимая связь.

Если значение статистики "хи-квадрат" меньше табличного критического значения, то наблюдаемые данные не сильно отличаются от тех, которых можно было бы ожидать, если бы переменные в генеральной совокупности были независимы. В этом случае следует принять нулевую гипотезу о независимости переменных (как приемлемую возможность) и сделать вывод, что между переменными отсутствует значимая связь. Это слабое заключение, поскольку принята нулевая гипотеза о независимости переменных: вы *признали* независимость переменных, но вы *не доказали* ее.

Основные термины

- Критерий "хи-квадрат" (chi-squared test), 879
- Статистика "хи-квадрат" (chi-squared statistic), 879
- Критерий "хи-квадрат" на равенство процентов (chi-squared test for equality of percentages), 881
- Независимый (independent), 888
- Условные проценты генеральной совокупности (conditional population percentages), 888
- Критерий "хи-квадрат" независимости (chi-squared test for independence), 889

Контрольные вопросы

1. Для каких переменных используют критерий "хи-квадрат"?
2. а) Что измеряет статистика "хи-квадрат" с точки зрения отношения между наблюдаемыми данными и нулевой гипотезой?

   б) В каком случае отклоняют нулевую гипотезу: при больших или при малых значениях статистики "хи-квадрат"? Почему?

3. Какова цель теста "хи-квадрат" на равенство процентов?
4. а) Для каких данных можно использовать критерий "хи-квадрат" на равенство процентов?

   б) Что представляют собой опорные значения в таком тесте?

   в) Сформулируйте соответствующие гипотезы.

   г) Как вычисляются ожидаемые частоты? Что они представляют?

   д) Какие допущения должны выполняться, чтобы можно было применять этот критерий?
5. а) Какой вывод можно сделать при тестировании равенства процентов, если значение статистики "хи-квадрат" *больше* табличного значения?

   б) Какой вывод можно сделать, если значение статистики "хи-квадрат" *меньше* табличного значения?
6. а) Что понимают под независимостью двух качественных переменных?

   б) Какова связь между условными вероятностями и независимостью качественных переменных?
7. Какова цель теста "хи-квадрат" независимости?
8. а) Для каких данных можно применять тест "хи-квадрат" независимости?

   б) Что представляют собой опорные значения (если они есть) в этом тесте?

   в) Сформулируйте соответствующие гипотезы.

   г) Как вычисляются ожидаемые частоты? Что они представляют?

   д) Какие допущения должны выполняться для того, чтобы можно было применять этот тест?
9. а) Какой вывод можно сделать при тестировании на независимость, если значение статистики "хи-квадрат" *больше* табличного значения?

   б) Какой вывод можно сделать, если значение статистики "хи-квадрат" *меньше* табличного значения?
10. Почему намного труднее установить независимость переменных, чем их зависимость (отсутствие независимости)?

Задачи

1. а) Если значение наблюдаемой частоты равно 3, а ожидаемой — 8,61, стоит ли продолжать выполнять тест "хи-квадрат"?

   б) Если значение наблюдаемой частоты равно 8, а ожидаемой — 3,29, стоит ли продолжать выполнять тест "хи-квадрат"?
2. Для каждого потенциального клиента, входящего в автосалон, зафиксирован предпочитаемый им тип автомобиля. В табл. 17.4.1 приведены данные за прошедшую неделю, а также соответствующие проценты в это же время в прошлом году.

   а) Вычислите проценты для данных за прошлую неделю.

Таблица 17.4.1. Предпочитаемый тип автомобиля

Тип	Частоты за прошлую неделю	Данные за прошлый год, %
Семейный седан	187	25,8
Малолитражный автомобиль	206	46,2
Спортивный автомобиль	29	8,1
Автомобиль фургон	72	12,4
Пикап	42	7,5
Итого	536	100,0%

   б) Сравните проценты данных за прошедшую неделю и данные за прошлый год. Опишите замеченные различия таким образом, чтобы это могло быть полезно продавцу автомобилей.

   в) Подсчитайте, сколько человек из 536 предпочло бы малолитражный автомобиль, если бы сохранилась тенденция прошлого года. Сравните это значение с наблюдаемым.

   г) Предполагая, что тенденции прошлого года сохраняются, вычислите ожидаемую частоту для каждого типа транспортных средств.

   д) Рассматривая проценты за прошлый год как точные, вычислите статистику "хи-квадрат".

   е) Обсудите допущения, необходимые для корректного применения критерия "хи-квадрат". В частности, что представляет собой генеральная совокупность, для которой вы делаете статистический вывод?

   ж) Чему равно число степеней свободы для этого теста "хи-квадрат"?

   з) Найдите соответствующие значения из таблицы "хи-квадрат" для уровней значимости 5%, 1% и 0,1%.

   и) Выполните тест "хи-квадрат" на каждом из этих уровней значимости и изложите полученные результаты.

   к) Сформулируйте выводы (с p-значением, указанным как $p > 0{,}05$, $p < 0{,}05$, $p < 0{,}01$ или $p < 0{,}001$) о каких-либо изменениях в предпочтениях покупателей.
3. В табл. 17.4.2 приведено процентное распределение количества входящих в вашу фирму телефонных звонков в это же время в прошлом году.

   а) Вычислите процентное распределение для количества звонков за первый день текущего месяца и сравните полученный результат с прошлогодним.

   б) Найдите ожидаемое количество звонков в первый день месяца при допущении, что прошлогодняя тенденция звонков сохранится и в этом месяце.

   в) Вычислите значение статистики "хи-квадрат" и число степеней свободы.

   г) Сформулируйте результат теста "хи-квадрат" на уровне значимости 5%.

   д) Сформулируйте вывод об изменении структуры тематики звонков в настоящее время по сравнению с тем же временем в прошлом году.

Таблица 17.4.2. Входящие телефонные звонки

Тип	Частота (первый день месяца)	Процент от общего количества (этот же месяц прошлого года), %
Заказ	53	33,2
Информация	54	38,1
Запрос на обслуживание	28	12,5
Отмена	18	9,7
Другое	7	6,5
Итого	160	100

4. Из выбранных случайным образом для проверки 267 роликовых коньков у пяти оказались потерянными заклепки, а 12 коньков технически не соответствовали спецификации.

   а) С помощью формулы для вычисления ожидаемой частоты определите, сколько роликовых коньков будут иметь *оба* этих дефекта при условии, что эти дефекты не зависят друг от друга?

   б) Используя относительную частоту, сделайте оценку вероятности дефекта отсутствия заклепки.

   в) Аналогичным образом вычислите оценку вероятности дефекта несоответствия спецификации.

   г) Используя формулу вычисления вероятности из главы 6, определите оценку вероятности первого и второго дефектов, предполагая независимость этих дефектов и используя значения оценок вероятностей из пп. "а" и "б".

   д) Преобразуйте полученное вами в п. "г" значение вероятности в ожидаемую частоту, умножив его на размер выборки.

   е) Сравните ваши результаты, полученные в п. "а" (из формулы для ожидаемых значений) и в п. "д" (из формулы независимости). Объясните, почему оба этих подхода дали такой результат.

5. Ваша фирма рассматривает вопрос расширения своего влияния на соседний город. При опросе служащим этого города был задан вопрос: "Как вы считаете, условия для бизнеса в вашем регионе стали лучше, хуже или остались теми же?" Результаты опроса приведены в табл. 17.4.3.

Таблица 17.4.3. Результаты опроса о перспективах бизнеса

	Менеджеры	Другие служащие	Итого
Лучше	23	185	
Те же	37	336	
Хуже	11	161	
Не знаю	15	87	
Итого			

а) Вычислите необходимые значения и заполните колонку и строку "Итого".

б) Постройте таблицу общих процентов. Интерпретируйте их как оценки вероятностей в генеральной совокупности. В частности, поясните, какие вероятности они представляют?

в) Постройте таблицу процентов по типам служащих. Интерпретируйте полученные значения как оценки вероятностей в генеральной совокупности. В частности, поясните, какие вероятности они представляют?

г) Постройте таблицу процентов по вариантам ответов. Интерпретируйте полученные значения как оценки вероятностей в генеральной совокупности. В частности, поясните, какие вероятности они представляют?

д) *Похоже ли*, что ответ не зависит от типа служащего? Почему да или почему нет?

6. Обратитесь к данным задачи 5.

   а) О чем утверждает (с точки зрения практики) нулевая гипотеза о независимости для этой ситуации?

   б) Как много менеджеров, которые ответили "Хуже", можно было бы ожидать в этой выборке, если бы ответ на это вопрос не зависел от типа служащего?

   в) Постройте таблицу ожидаемых частот, предполагая независимость переменных.

   г) Вычислите статистику "хи-квадрат".

   д) Определите число степеней свободы для этого теста "хи-квадрат".

7. Обратитесь к данным задачи 5.

   а) Найдите в таблице "хи-квадрат" критическое значение для уровня значимости 5% и изложите результат теста "хи-квадрат".

   б) Найдите в таблице "хи-квадрат" критическое значение для уровня значимости 1% и изложите результат теста "хи-квадрат".

   в) Найдите в таблице "хи-квадрат" критическое значение для уровня значимости 0,1% и изложите результат теста "хи-квадрат".

   г) Сформулируйте заключение (с p-значением, указанным как $p > 0,05$, $p < 0,05$, $p < 0,01$ или $p < 0,001$) и поясните результат с практической точки зрения.

8. Рассмотрим представленные в табл. 17.4.4 результаты небольшого опроса о возможности в ближайшие двенадцать месяцев краха фондового рынка, сравнимого с крахом, произошедшим в 1987 году.

   а) Вычислите необходимые значения и заполните колонку и строку "Итого".

   б) Постройте таблицу общих процентов. Интерпретируйте их как оценки вероятностей в генеральной совокупности. В частности, поясните, какие вероятности они представляют?

   в) Постройте таблицу процентов по типам респондентов (акционер/не акционер). Проинтерпретируйте полученные значения как оценки вероятно-

Таблица 17.4.4. Ответы на вопрос о возможном большом крахе фондового рынка

	Акционеры	Не акционеры	Итого
Очень вероятно	18	26	
Весьма вероятно	41	65	
Маловероятно	52	68	
Невероятно	19	31	
Не уверен	8	13	
Итого			

стей в генеральной совокупности. В частности, поясните, какие вероятности они представляют?

г) Постройте таблицу процентов по вариантам ответа. Интерпретируйте полученные значения как оценки вероятностей в генеральной совокупности. В частности, поясните, какие вероятности они представляют?

д) *Похоже ли*, что ответы не зависят от классификации акционер/не акционер? Почему да или почему нет?

9. Обратитесь к данным задачи 8.

а) Что утверждает применительно к практической задаче нулевая гипотеза о независимости?

б) Как много акционеров, которые ответили "Очень вероятно", можно было бы ожидать в этой выборке, если бы ответ не зависел от классификации акционер/не акционер?

в) Предполагая независимость переменных, постройте таблицу ожидаемых частот.

г) Вычислите статистику "хи-квадрат".

д) Определите число степеней свободы для теста "хи-квадрат".

10. Обратитесь к данным задачи 8.

а) Найдите в таблице "хи-квадрат" критическое значение для уровня значимости 5% и изложите результат теста "хи-квадрат".

б) Найдите в таблице "хи-квадрат" критическое значение для уровня значимости 1% и изложите результат теста "хи-квадрат".

в) Найдите в таблице "хи-квадрат" критическое значение для уровня значимости 0,1% и изложите результат теста "хи-квадрат".

г) Сформулируйте заключение (с p-значением, указанным как $p > 0{,}05$, $p < 0{,}05$, $p < 0{,}01$ или $p < 0{,}001$) и поясните результат с практической точки зрения.

11. Компанию, доставляющую товары почтой, интересует, одинаков ли уровень заказов (процент отправленных по почте каталогов, которые стали основой заказа) в разных регионах. В табл. 17.4.5 приведены последние данные по регионам о количестве отправленных по почте каталогов, которые стали причиной заказа и которые не стали причиной заказа.

Таблица 17.4.5. Количество заказов по регионам

	Восток	Запад
Заказ сделан	926	352
Заказ не сделан	22 113	10 617

а) Определите процент заказов в каждом из регионов. В каком регионе этот показатель выше?

б) Значимо ли различаются проценты заказов в этих регионах? Как вы это определили?

12. Коммерческий банк анализирует состояние недавних заявок на получение ссуды под залог недвижимости. Некоторые заявки приняты, некоторые отклонены, а некоторые временно отложены до получения новой информации. Полученная информация представлена в табл. 17.4.6 и на рис. 17.4.1.

а) Представьте себя инспектором и исходя из рис. 17.4.1 напишите небольшой отчет, в котором сравните состояние заявок на получение ссуды под залог жилых и промышленных зданий.

б) Значима ли разница между клиентами, работающими с жилыми зданиями, и клиентами, работающими с промышленными зданиями? Как вы это определили?

13. Имеет ли значение, как вы задаете вопрос? Было проведено исследование, в ходе которого испытуемым задавали вопрос о том, согласны ли они за-

Таблица 17.4.6. Состояние заявок на ссуду под залог недвижимости

	Жилые здания	Промышленные здания
Принято	78	57
Требуется информация	30	6
Отклонено	44	13

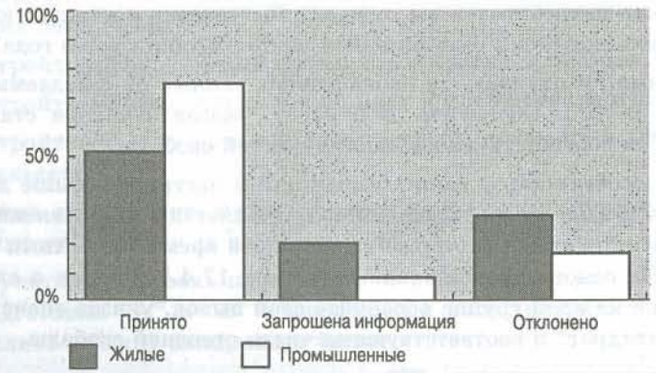


Рис. 17.4.1. Распределение состояния заявлений на получение ссуды под залог жилых и промышленных зданий

платить 30 долларов за обед в конкретном ресторане?[6] При этом одной группе говорили, что "с вероятностью 50% вы будете довольны", а другой говорили, что "с вероятностью 50% вы будете недовольны". Единственная разница в формулировке состояла в том, что в одном случае использовали слово "довольны", а в другом слово " недовольны". В результате опроса выяснилось, что 26% из 240 человек, которым был задан вопрос со словом "довольны", ответили, что они будут посещать этот ресторан, по сравнению с 11% из 215, которым был задан вопрос со словом "недовольны". Значима ли эта разница (26% и 11%), или разницу такого размера можно получить только лишь за счет случайности? Как вы это определили?

14. На восточной фабрике зафиксировано 28 несчастных случаев при численности рабочей силы 673 человека. За этот же период времени на западной фабрике зафиксирован 31 несчастный случай при численности рабочей силы 1306 человек.

    а) На какой фабрике больше несчастных случаев? На какой фабрике выше уровень несчастных случаев?

    б) Есть ли значимое различие между уровнями несчастных случаев на этих двух фабриках? Обоснуйте ваш ответ, указав значение статистики "хи-квадрат" и соответствующее число степеней свободы.

15. При опросе в одной группе домохозяйств был задан вопрос о том, насколько семья удовлетворена своим автомобилем, а в другой группе домохозяйств был задан вопрос о том, насколько семья не удовлетворена своим автомобилем. Полученные данные приведены в табл. 17.4.7.

    а) В какой группе вероятность ответа "удовлетворяет" была выше?

    б) В какой группе вероятность ответа "не удовлетворяет" была выше?

    в) Значима ли эта разница? Обоснуйте ваш ответ, указав значение статистики "хи-квадрат" и соответствующее число степеней свободы.

16. Ниже приведены поквартальные данные прошедшего года о количестве новых потребителей, которые подписались на полный перечень услуг: 106, 108, 72 и 89.

    а) Сколько потребителей вы ожидали бы увидеть в каждом квартале, если бы уровень подписки был одинаков на протяжении всего года?

    б) Значимо ли отличаются наблюдаемые данные от ожидаемых (в соответствии с п. "а"). Обоснуйте свой ответ, указав значение статистики "хи-квадрат" и соответствующее число степеней свободы.

17. Есть ли особенность у ваших клиентов? В частности, выше ли уровень их интереса к вашему информационному бюллетеню по сравнению с потенциальными клиентами (которые в настоящее время клиентами еще не являются)? На основе представленных в табл. 17.4.8 данных о случайных выборках по каждой группе обоснуйте свой вывод, указав значение статистики "хи-квадрат" и соответствующее число степеней свободы.

[6] Peterson R. A. and Wilson W. R. "Measuring Customer Satisfaction: Fact and Artifact". *Journal of the Academy of Marketing Science*, 1992, 20, p. 61-71.

Таблица 17.4.7. Ответы членов домохозяйств в зависимости от формулировки вопроса

	Формулировка вопроса	
	"удовлетворяет"	"не удовлетворяет"
Полностью удовлетворяет	139	128
Почти удовлетворяет	82	69
Частично не удовлетворяет	12	20
Полностью не удовлетворяет	10	23

Данные предоставлены Peterson R. A. and Wilson W. R. "Measuring Customer Satisfaction: Fact and Artifact", *Journal of the Academy of Marketing Science*, 1992, 20, p. 61–71.

Таблица 17.4.8. Интерес к информационному бюллетеню клиентов и потенциальных клиентов

	Клиент	Потенциальный клиент
Очень интересует	49	187
Немного интересует	97	244
Не интересует	161	452

Упражнения с использованием базы данных

Обратитесь к базе данных служащих, приведенной в приложении А.

1. Можно ли сказать, если исключить случайность, что приблизительно одинаковое количество служащих имеют уровни подготовки "А", "В" и "С"? (Чтобы ответить на этот вопрос, проверьте, значимо ли отличается процент служащих каждого уровня подготовки от пропорций 1/3, 1/3 и 1/3 для этих трех уровней.)
2. Есть ли свидетельства дискриминации по признаку пола в отношении уровня подготовки? Чтобы ответить на этот вопрос, выполните следующее.

    а) Постройте таблицу частот для двух качественных переменных: "пол" и "уровень подготовки".

    б) Постройте таблицу общих процентов и кратко прокомментируйте ее.

    в) Постройте таблицу процентов по полу и прокомментируйте результаты.

    г) Постройте таблицу процентов по уровню подготовки и прокомментируйте результаты.

    д) Можно ли для этого набора данных использовать тест "хи-квадрат" независимости? Почему да или почему нет?

    е) Не учитывая уровень подготовки "С", ограничьте рассмотрение уровнями подготовки "А" и "В" и продолжите выполнение этого упражнения. Вычислите таблицу ожидаемых частот.

    ж) Не учитывая уровень подготовки "С", вычислите значение статистики "хи-квадрат".

з) Не учитывая уровень подготовки "С", сформулируйте результат теста "хи-квадрат" независимости для уровня значимости 5%. Прокомментируйте полученные результаты.

Проекты

1. Найдите в Internet, в газетах или журналах данные об одной качественной переменной, имеющей отношение к вашей работе или вашим бизнес-интересам, а также перечень опорных значений процентов для сравнения.



а) Постройте на основе имеющихся наблюдений таблицу частот.

б) Постройте на основе имеющихся наблюдений таблицу процентов.

в) Сравните полученные вами проценты с опорными и прокомментируйте сходство и различие.

г) Постройте таблицу ожидаемых частот в предположении, что опорные значения процентов справедливы для рассматриваемой генеральной совокупности.

д) Для каждой категории укажите

$$\left[(\text{Наблюдаемое значение} - \text{Ожидаемое значение})^2 / \text{Ожидаемое значение}\right].$$

Найдите в этом перечне наибольшее и наименьшее значение и объясните, почему эти конкретные категории являются наименьшими и наибольшими, сравнивая наблюдаемые и ожидаемые проценты для этих же категорий.

е) Вычислите значение статистики "хи-квадрат".

ж) Выполните тест "хи-квадрат" и определите p-значение (представив результат как $p > 0,05$, $p < 0,05$, $p < 0,01$ или $p < 0,001$).

з) Прокомментируйте, о чем свидетельствует тест "хи-квадрат" с точки зрения рассматриваемой экономической ситуации.

2. Найдите в Internet, в газетах или журналах данные о двух качественных переменных (двумерный набор данных), связанных с вашей работой или с вашими интересами в бизнесе.



а) Постройте из полученных наблюдений таблицу частот.

б) Постройте на основе этих наблюдений таблицу общих процентов и прокомментируйте ее

в) Постройте на основе этих наблюдений таблицу процентов по одной из переменных. Дайте комментарий с точки зрения профиля другой переменной.

г) Повторите предыдущий пункт, используя проценты по другой переменной.

д) Сформулируйте для этой ситуации нулевую гипотезу независимости переменных. Представляет ли она собой приемлемую возможность, с вашей точки зрения?

е) Постройте таблицу ожидаемых частот, предполагая независимость этих двух переменных.

ж) Укажите для каждой комбинации категорий значение

$$\left[(\text{Наблюдаемое значение} - \text{Ожидаемое значение})^2 / \text{Ожидаемое значение}\right].$$

Найдите в этом перечне наибольшее и наименьшее значение и объясните, почему эти конкретные комбинации категорий являются самыми маленькими и самыми большими, сравнивая наблюдаемые и ожидаемые проценты для этих категорий.

з) Вычислите значение статистики "хи-квадрат".

и) Выполните тест "хи-квадрат" и определите p-значение (представив результат как $p > 0,05$, $p < 0,05$, $p < 0,01$ или $p < 0,001$).

к) Объясните, о чем свидетельствует тест "хи-квадрат" в отношении рассматриваемой экономической ситуации.

ГЛАВА 18

Контроль качества: выявление вариации и управление ею

Статистический контроль качества заключается в применении статистических методов для оценки и улучшения результатов любой деятельности. Экономическую деятельность, относящуюся к производству или сфере услуг, можно контролировать, чтобы иметь возможность выявить и откорректировать проблемы прежде, чем они серьезно осложнят ситуацию. Для решения таких задач хорошо подходят статистические методы, поскольку разумные управленческие решения должны быть основаны, по крайней мере частично, на *данных*. Ниже приведено несколько примеров, иллюстрирующих разнообразные ситуации, в которых можно использовать статистический контроль качества для укрепления позиции фирмы.



Пример 1. Фирме очень не нравится, когда потребитель возвращает купленный товар. Кроме потери выручки за продажу, неприятно также то, что потребитель может изменить в худшую сторону мнение о вашей фирме. Почему бы не рассмотреть эту проблему с точки зрения возможности улучшения работы? Собрав данные, касающиеся различных причин возврата товара, вы получите богатую информацию, которую можно использовать для разных целей. Анализ этого перечня причин позволит сконцентрировать ваше внимание на основных проблемах и улучшить качество продукта. Кроме того, вы больше узнаете о потребителях. Такая собственная база данных может быть полезна в маркетинге и коммерческой деятельности, а также при разработке новых изделий.

Пример 2. На упаковке написано, что каждая пачка средства для мытья посуды весит 16 унций. Если бы проверка веса стоила недорого, было бы неплохо убедиться, что каждая пачка весит *точно* 16 унций. Однако расходы на проверку слишком велики. Некоторый уровень колебания веса пачек допустим, и вы хотите проконтролировать этот уровень. Одна из целей состоит в том, чтобы избежать неприятностей, касающихся связей с общественностью. Для этого нужно быть уверенным, что ни в одной из пачек нет существенного недовеса и, по крайней мере, средний вес пачки составляет 16 унций. Другая цель состоит в том, чтобы снизить расходы за счет недопущения слишком сильного превышения веса. Проанализировав чистый вес этих пачек (либо каждой пачки, либо случайную выборку из каждой партии), можно получить информацию о производственном процессе и уровне его вариации. Когда станет ясно, что процесс управляем, вес можно уже не контролировать. Подходящий метод контроля может даже позволить установить проблему и определить тренд, прежде чем продолжающееся отклонение процесса приведет к серьезным неприятностям.

Пример 3. Ваша бухгалтерия выполняет важную функцию: превращает денежные поступления фирмы в наличные средства. Любые проблемы с регулярностью поступления средств в бухгалтерию переходят непосредственно в потерю стоимости, например потерю наличности, которую можно было бы получить, если бы картотеку не держали три недели в ожидании внутреннего решения. И не забывайте, что запаздывание получения наличности приводит к снижению ее стоимости. Сделав выборку из данных о проведении расчетов, которые вызывают и не вызывают затруднения, можно установить, управляем ли процесс погашения дебиторской задолженности. Проанализировав повторяющиеся проблемы, можно определить наиболее уязвимые части системы, которые нуждаются в изменении. Возможно, необходимо увеличить количество сотрудников. Возможно, некоторые виды работ могут выполняться "параллельно", чтобы одна группа не ждала, пока другая закончит свою работу.

Контроль качества хорошо рассматривать на шкале соотношения затрат и прибыли. Расходы на осуществление программы статистического контроля качества обычно невелики по сравнению с денежными средствами, сэкономленными в результате ее реализации. Когда выпуск некачественных изделий устранен, вы экономите, потому что нет необходимости в проверке каждого изделия. Наконец, репутация высокого качества и стабильности будет способствовать заключению контрактов по благоприятным ценам.

Но не думайте, что статистические методы все сделают за вас. Они могут только предоставить информацию, а ваша задача — наилучшим образом воспользоваться ею. К примеру, статистические методы могут предоставить вам информацию о том, что, вероятно, что-то произошло около 10:30, потому что упаковки в это время стали значительно тяжелее, но необходимая регулировка оборудования по-прежнему зависит только от вас и ваших рабочих.

Не ждите невозможного от статистических методов. Возможности системы сами по себе также должны быть учтены. Если сверло настолько старо и изношено, что им невозможно просверлить ровное отверстие, то никакие статистические методы не решат эту проблему. Хотя хорошая программа контроля качества поможет вам получить максимум из имеющегося оборудования, вы можете

обнаружить, что необходима определенная модернизация, прежде чем будут достигнуты приемлемые результаты.

Все четыре основные стадии статистического анализа данных играют важную роль в контроле качества. Стадия *планирования* включает определение конкретных процессов, требующих проверки, и методов измерения параметров этих процессов. На остальных трех стадиях для представления данных часто используют *карты контроля*. Стадия *исследования* включает рассмотрение данных с точки зрения конкретной проблемы в изучаемом процессе. Стадия *оценки* включает отображение текущего состояния процесса и того, насколько хороши результаты функционирования. На стадии *проверки гипотез* принимают решение о том, необходимо вмешиваться в процесс или нет.

В 1950-е годы В. Эдвардс Деминг (W. Edwards Deming) применил статистические методы контроля качества в Японии и в дальнейшем продолжал оказывать помощь фирмам во всем мире при реализации их программ контроля качества. Ниже приведены 14 пунктов рекомендаций Деминга для управления процессом улучшения качества[1].

1. *Сделайте улучшение качества продукции и услуг постоянной задачей*, поставив себе цель быть конкурентоспособным, утвердиться в бизнесе и создавать рабочие места.
2. *Примите новый образ мышления (философию)*. Мы находимся в новом экономическом веке, созданном Японией. Нельзя больше мириться ни с общепринятым в Америке стилем управления, ни с общепринятыми уровнями задержек, ошибок или некачественных изделий.
3. *Положите конец зависимости качества от контрольных проверок*. Исключите необходимость массового контроля, поставив на первое место качество изготовления изделия.
4. *Прекратите строить свой бизнес только исходя из цены товара, указанной на этикетке*. Вместо этого минимизируйте общие расходы.
5. *Постоянно улучшайте систему производства продукции и услуг, чтобы повысить качество и производительность*, и таким образом постоянно снижайте издержки.
6. *Сделайте обучение частью работы*.
7. *Введите контроль*, целью которого должна быть помощь людям и машинам наилучшим образом выполнять свои функции. Необходимо пересмотреть как контроль менеджмента, так и контроль производственных рабочих.
8. *Избавьтесь от сомнений*: каждый способен работать эффективно.
9. *Устраните барьеры между отделами*. Исследователи, конструкторы, производственники и коммерческие агенты должны работать как одна команда, чтобы предвидеть проблемы, которые могут встретиться при производстве изделий или услуг.

[1] Эти 14 пунктов перепечатаны из *The ESB Journal*, Spring 1989 (Educational Service Bureau of Dow Jones & Co., Inc.). Дальнейшее обсуждение этих пунктов можно найти в книгах Edwards Deming W. *Out of the Crisis* (Cambridge, Mass.: MIT Center for Advanced Engineering Studies, 1986) и в Gitlow H. A. Oppenheim, and R Oppenheim, *Quality Management: Tools and Methods for Improvement*, 2nd ed. (Burr Ridge, III.: Richard D. Irwin, 1995).

10. *Исключите лозунги и плакаты, которые призывают сотрудников работать без брака и на новом уровне производительности*. Такие призывы только создают взаимоотношения соперничества. Основные причины низкой производительности кроются в системе, которая находится за пределами полномочий работников.
11. *Исключите ежедневные количественные нормативы*. Вместо них используйте вспомогательные средства и разумный контроль.
12. *Удалите барьеры, которые лишают обычного рабочего возможности гордиться своим мастерством*. Контролировать следует не количество, а качество. Это также означает отмену ежегодного рейтинга или оценки квалификации, а также управление с помощью цели.
13. *Введите сильную программу обучения и подготовки*.
14. *Добейтесь, чтобы каждый сотрудник компании работал над совершенствованием преобразований*.

18.1. Процессы и причины вариации

Под **процессом** понимают любой вид экономической деятельности, который превращает ресурсы в продукты. Производственный процесс — это процесс, посредством которого сырье превращают в изделия. В ресторанах осуществляются процессы, которые преобразуют продукты и энергию в готовые к употреблению блюда. В офисах имеют место различные процессы, которые преобразуют информацию из одного вида (например, некоторые записанные на бумаге документы) в другой (например, компьютеризованные отчеты или чек для оплаты).

Процесс может включать другие процессы, называемые *подпроцессами*, каждый из которых также является процессом. Например, построение самолета можно рассматривать как один процесс, который использует различные ресурсы (металл, пластмассу, провод, хранящуюся в компьютере информацию) и превращает их в самолеты. Это сложный процесс, состоящий из множества подпроцессов (например, сборка фюзеляжа, монтаж освещения кабины, проверка работы закрылков). Каждый из этих подпроцессов, в свою очередь, состоит из подпроцессов (например, вставка одной заклепки, пайка одного провода, проверка максимального натяжения троса). Исходя из поставленной цели можно сфокусировать свое внимание на необходимом уровне детальности. Методы статистического контроля процессов можно применять к любому процессу или подпроцессу.

Статистический контроль процесса заключается в использовании статистических методов для наблюдения за функционированием процесса таким образом, чтобы можно было отрегулировать процесс, при необходимости устранить неполадки или не вмешиваться в процесс, если он протекает должным образом. Цель статистического контроля состоит в обнаружении проблем и их устранении *до того*, как будут выпущены бракованные изделия. Использование выборочных данных для статистического контроля позволяет обеспечить выпуск высококачественной продукции, не проверяя каждую отдельную единицу!

Почти у каждого процесса показатели подвержены вариации. Некоторые изменения показателей малы и несущественны, как, например, точное количество

стружек шоколада, которое незначительно изменяется у разных кондитеров. Другие изменения играют важную роль, как, например, деформированная металлическая коробка, похожая на современную абстрактную скульптуру, но не на то, что необходимо потребителю.

Разница между тем, "что хотелось" и "что получилось", может быть обусловлена любыми причинами. Одни причины легко выявить, другие — сложнее. Иногда причину можно найти даже прежде, чем будут получены данные (например, обнаружить дым по запаху). Некоторые причины требуют расследования, чтобы выявить, что является источником (скажем, тот станок, который необходимо отрегулировать) и почему (вероятно, служащему необходимы новые очки). Некоторые отклонения не стоят даже усилий на их исследование (как, например, почему при производстве ленты в коробку иногда попадает на дюйм ленты больше, что приводит к очень малому перерасходу сырья на выпуск этой коробки).

Всякий раз, когда можно приемлемо выяснить, почему возникла та или иная проблема, имеет место **неслучайная причина отклонений**. Обратите внимание, что фактически вы можете не знать действительную причину; достаточно того, что вы смогли ее обнаружить без привлечения больших расходов. Ниже приведены примеры таких неслучайных причин и возможные пути их нейтрализации.

1. В "чистую комнату" попадает пыль и мешает производству микрокристаллов и дисководов. Возможные решения включают контроль очистительного оборудования, при необходимости замену фильтров и изоляции, а также пересмотр процедуры входа и выхода служащих из помещения.
2. Клерки неправильно заполняют бланки форм, записывая вычисленные суммы не в те клетки таблиц. Возможные решения включают обучение клерков, изменение бланков или оба мероприятия одновременно.

Все причины отклонений, на выявление которых не стоит тратить усилий, рассматривают вместе как **случайные причины отклонений**[2]. Эти причины следует рассматривать как такие, которые определяют базовую случайность ситуации таким образом, что вы знаете, каким будет результат, если ситуация находится под контролем, и не ожидаете слишком многого. Ниже приведены примеры отклонений, обусловленных случайными причинами.

1. Механизм по заполнению бутылок содовой отрегулирован с точностью до капли, но несмотря на это, объем жидкости несколько отличается в разных бутылках.
2. Количество кукурузных хлопьев не одинаково во всех коробках. Нет необходимости контролировать эти колебания при условии, что они находятся в пределах допустимого и общий вес пачки практически соответствует весу, указанному на упаковке.

[2] Это не совсем строго соответствует принятому в статистике определению слова *случайный*. См., например, определение случайной выборки в главе 8. Однако отдел статистики Американского общества контроля качества в *Glossary and Tables for Statistical Quality Control*, p. 29 (Milwaukee, Wis.: American Society for Quality Control, 1983) так определяет *случайные причины*: "Факторы, обычно многочисленные и по отдельности относительно мало значимые, которые влияют на вариацию, но которые практически невозможно обнаружить или выделить".

Если все неслучайные причины колебаний определены и исключены, так что остались только случайные причины, то говорят, что процесс находится **в состоянии статистического контроля**, или **под контролем**. Если процесс находится под контролем, то за ним достаточно только наблюдать. Если процесс *выходит из-под контроля*, то это означает появление неслучайной причины и необходимость решать возникшую проблему.

Программы контроля качества больше не являются только внутренним делом компании. Все больше фирм требуют от поставщиков подтверждения (с предоставлением карты контроля), что продаваемая ими продукция произведена в контролируемых условиях. Если у вас есть проблемы с поставщиком, вы можете рассмотреть и такую возможность.

Диаграмма Парето показывает, на что обратить внимание

Предположим, что вы изучили группу некачественных компонентов и классифицировали каждую единицу в соответствии с причиной дефекта. **Диаграмма Парето** показывает причины различных дефектов, упорядоченные от наиболее к наименее часто встречающимся, что позволяет сосредоточить внимание на самых важных проблемах. Кроме количества и процента дефектов, обусловленных каждой из причин, диаграмма показывает также *кумулятивный процент*, что позволяет легко определить процент дефектов, вызванных двумя (тремя, четырьмя) наиболее важными причинами одновременно.

Ниже описана последовательность шагов при построении диаграммы Парето.

1. Начните с количества дефектов (частоты) для каждой из причин. Определите общее количество дефектов и процент дефектов, обусловленных каждой из причин.
2. Расположите причины в порядке уменьшения частоты дефектов, т.е. таким образом, чтобы более важные проблемы стояли в начале списка.
3. Постройте столбиковую диаграмму для этих частот.
4. Для каждой причины дефекта вычислите кумулятивный процент путем сложения процента, соответствующего этой причине, со всеми процентами, расположенными в списке перед этой причиной.
5. Начертите линию (выше столбиков на диаграмме), соответствующую этим кумулятивным процентам.
6. Расставьте на диаграмме названия. Под каждым столбиком на диаграмме укажите название соответствующей причины. Слева на вертикальной оси укажите количество дефектов. На вертикальной оси справа укажите *кумулятивные проценты* дефектов.

Например, рассмотрим проблемы, которые возникают в процессе производства небольших устройств бытовой электроники (табл. 18.1.1).

Набор этих же данных, но упорядоченных таким образом, что причины расположены по убыванию вычисленных процентов, показан в табл. 18.1.2.

Представленная на рис. 18.1.1 диаграмма Парето показывает, что наибольшую проблему представляет блок электропитания, поскольку с ним связано вдвое больше количество дефектов, чем со следующей по важности проблемой производства пластиковой коробки. Линия кумулятивных процентов демонстрирует, что с этими

двумя наиболее важными проблемами связана основная (85,9%) часть всех дефектов. С тремя наиболее важными проблемами связано почти 97,2% всех дефектов.

Таблица 18.1.1. Причины дефектов и частоты их появления

Причина проблемы	Количество случаев
Пайка соединений	37
Пластиковая коробка	86
Блок электропитания	194
Грязь	8
Удар (изделие уронили)	1
Итого	326

Таблица 18.1.2. Упорядоченные по убыванию частот причины с указанием соответствующих процентов и кумулятивных процентов

Причина проблемы	Количество случаев	Процент	Кумулятивный процент
Блок электропитания	194	59,5	59,5
Пластиковая коробка	86	26,4	85,9
Пайка соединений	37	11,3	97,2
Грязь	8	2,5	99,7
Удар (изделие уронили)	1	0,3	100,0
Итого	326	100,0%	

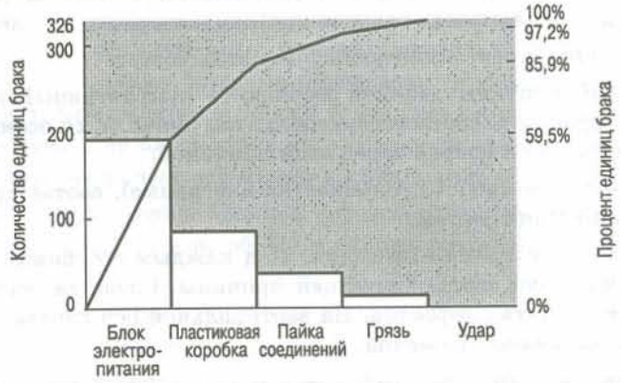


Рис. 18.1.1. Диаграмма Парето для причин дефектов при производстве небольших устройств бытовой электроники. Высота столбика показывает важность (частоту дефектов) каждой из причин. Например, 86 или 26,4% всех дефектов связаны с пластиковой коробкой. Линия отражает кумулятивный процент основных причин брака. Например, две первые причины (проблемы блока электропитания и пластиковой коробки) вместе обусловливают 85,9% дефектов

На диаграмме Парето самый высокий столбик будет всегда слева (*показывая наиболее часто возникающие проблемы*), а самый короткий — справа. Линия, отражающая кумулятивный процент, всегда направлена вверх и имеет тенденцию приближаться к горизонтальной при движении вправо.

Полезная функция диаграммы Парето состоит в том, что она вносит некоторую объективность в споры о качестве. Вместо того чтобы дать служащим возможность заняться решением тех проблем, которые им наиболее знакомы и с которыми им больше нравится работать, диаграмма Парето помогает сконцентрировать их внимание на проблемах, наиболее важных для фирмы.

18.2. Что такое карты контроля и как их читать

Как только вы выбрали один из многих процессов, за которые вы отвечаете как менеджер, и один из многих методов измерения параметров этого процесса, вы хотите осмыслить эту информацию, чтобы узнать, когда нужно действовать, а когда *не* нужно. **Карта контроля** показывает последовательность замеров параметров процесса, *центральную линию* и *контрольные границы*, которые вычисляют, чтобы решить, находится процесс под контролем или нет. Если принимается решение о том, что процесс вышел из-под контроля, то карта контроля позволяет определить проблему, чтобы ее можно было решить.

Использование карт контроля для контроля качества продукции включает все четыре основные стадии статистического анализа. На стадии *проектирования* происходит выбор процесса и тех его параметров, которые необходимо измерять для составления карты контроля. За составлением карты контроля следуют еще три стадии. На стадии *исследования* изучают карту контроля, чтобы определить модели, тренды и исключительные случаи, которые характеризуют текущее состояние процесса и позволяют судить о вероятном развитии событий в будущем. Стадия *оценки* включает вычисление обобщающих параметров процесса, некоторые из которых приведены в карте контроля. И, наконец, на стадии *проверки гипотезы*, используя данные (результаты измерений), принимают решение о том, находится процесс под контролем или нет. Ниже приведены формулировки проверяемых гипотез.

H_0 : процесс находится под контролем.

H_1 : процесс не находится под контролем.

Обратите внимание, что нулевая гипотеза представляет собой предположение о том, что процесс находится под контролем. Утверждая так, вы гарантируете, что процесс не нуждается в регулировке до тех пор, пока не будет получено убедительное доказательство того, что возникла проблема. Регулировка процесса может стоить очень дорого, что обусловлено потерей производственного времени, расходами на саму регулировку, а также возможностью того, что в результате регулировки отклонения параметров системы могут *увеличиться*. Не стоит регулировать процесс, пока в этом нет необходимости. В таких случаях обычно говорят: "Если не сломано — не трогай !".

Уровень ложной тревоги характеризует то, как часто процесс регулируют без необходимости; это понятие аналогично понятию ошибки I рода в проверке гипотез. При построении карт контроля, как правило, используют интервал шириной в три, а не две, стандартные ошибки параметров, поскольку обычный для проверки статистических гипотез уровень ошибки I рода 5% недопустимо велик для большинства приложений контроля качества[3].

В остальных разделах этой главы рассказывается, как читать типичные карты контроля. Подробности составления различных типов карт изложены в разделе 18.3.

Контрольные границы показывают выход из-под контроля одного наблюдения

Если процесс находится под контролем, то мы ожидаем увидеть кривую, которая находится в пределах контрольных границ, как это показано на рис. 18.2.1.



Рис. 18.2.1. Эта карта контроля показывает процесс, который находится под контролем. Колебания результатов измерений случайны и не выходят за пределы контрольных границ

Все значения наблюдений находятся внутри контрольных границ. Хотя наблюдаемые значения различаются у разных единиц продукции, нет никакой закономерности этих различий. Процесс колеблется случайным образом внутри контрольных границ.

Если какое-либо значение выходит за пределы контрольных границ, либо выше верхней границы, либо ниже нижней границы, это является свидетельством того, что процесс вышел из-под контроля. Пример такого процесса показан на рис. 18.2.2.

С точки зрения статистики это означает, что следует отклонить нулевую гипотезу о том, что процесс контролируется, и принять исследуемую (альтернативную) гипотезу о том, что процесс не находится под контролем. Говоря практическим языком, возникает проблема, и карта контроля поможет нам двигаться в нужном для решения этой проблемы направлении. Карта показывает, какая именно единица продукции не соответствует норме, следовательно, можно просмотреть записи, относящиеся к этой единице, и поговорить с лицами, ответственными за ее выпуск. Можно быстро определить явную неслучайную причину (например, сломанное сверло или перегретую ванну для припоя). С другой стороны, может оказаться, что для установления причины дефекта необходимы дальнейшие исследования и проверки. А возможно, это редкий сигнал ложной тревоги, поэтому никакие корректирующие действия принимать не следует.

[3] Из главы 10 о проверке статистических гипотез мы знаем, что табличное t-значение для двусторонней проверки на уровне 5% равно 1,96, или приблизительно 2. Используя вместо 1,96 число 3, мы снижаем для больших выборок уровень ложной тревоги с 5% до 0,27%.

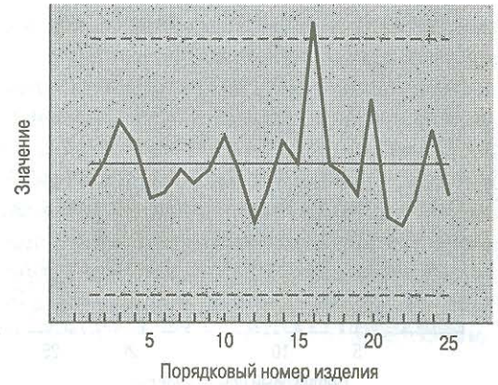


Рис. 18.2.2. Процесс не находится под контролем. Результат 16-го измерения выходит за пределы контрольных границ (находится выше верхней границы). Изучение обстоятельств производства этой отдельной единицы продукции поможет в будущем избежать дефектов такого вида

Как выявить проблему даже в пределах контрольных границ

Один из методов определения того, вышел ли процесс из-под контроля, заключается в анализе точек, которые вышли за пределы контрольных границ. Однако, даже если все точки находятся в пределах контрольных границ, исходя из карты контроля иногда становится ясно, что процесс вышел из-под контроля.

Идея заключается в том, чтобы рассмотреть структуру комбинаций точек в пределах контрольных границ. Даже если все точки находятся в пределах контрольных границ, но подозрительно близко сгруппировались возле одной из них (рис. 18.2.3) или целенаправленно движутся в направлении одной из границ (рис. 18.2.4), можно сделать вывод о том, что процесс больше не находится под контролем[4].

[4] Набор правил для вынесения решения о том, находится ли процесс под контролем, приведен в уже цитированной ранее книге Gitlow. В частности, можно сделать вывод о том, что процесс не находится под контролем, если восемь или более последовательных точек расположены по одну сторону от центральной линии или восемь или более точек образуют последовательность, направленную либо вверх, либо вниз.

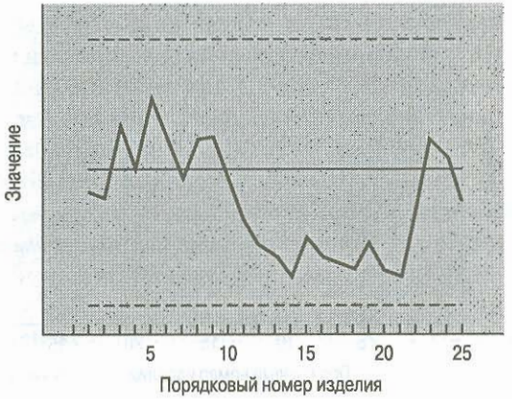


Рис. 18.2.3. Несмотря на то что все точки находятся в пределах контрольных границ, характер их расположения вызывает беспокойство. Обратите внимание на последовательность одиннадцати точек (с номерами от 12 до 22), которые находятся внутри границ, но близко подошли к нижней контрольной границе. Рассматривая этот факт в контексте всей карты, можно сделать вывод, что процесс не находится под контролем



Рис. 18.2.4. Все точки находятся в пределах контрольных границ, но, тем не менее, образуют тренд, который вызывает беспокойство. Обратите внимание, что последовательность точек устойчиво смещается вверх. Нельзя ждать до тех пор, пока точки выйдут за пределы верхней контрольной границы. Следует сделать вывод, что процесс не находится под контролем

Внезапный скачок значений на другой уровень карты контроля (см. рис. 18.2.3) позволяет сделать предположение о неожиданном изменении процесса. Возможно, что-то попало в привод или, может быть, работу принял новый сотрудник, еще не успевший достаточно хорошо изучить систему.

Постепенное смещение точек вверх или вниз (см. рис. 18.2.4) может быть следствием износа оборудования. Возможно, износились и начали медленно разрушаться отдельные части производственной линии, а может быть, кончился срок эксплуатации химической ванны и ее необходимо заменить.

Интерпретация карты контроля похожа на работу детектива. Изучая данные, тренды и структуру расположения точек, можно сделать вывод о том, на что следует обратить внимание, чтобы устранить возникшую проблему и вернуть процесс под контроль.

18.3. Отображение количественных измерений в $\bar{X}$- и R-картах

Для отображения количественных измерений обычно выбирают выборку относительно небольшого размера ($n = 4$ или $n = 5$) и затем наносят на карту результаты из многих последовательных выборок. Каждая точка на карте контроля отражает определенный показатель (центральное значение, или изменчивость) для n отдельных наблюдений. Рекомендуется выбрать небольшое значение n, поскольку процесс может меняться быстро, а вам желательно определить это изменение в процессе прежде, чем оно приведет к многочисленным дефектам.

Для отображения количественных измерений чаще всего используют $\bar{X}$- и R-карты. На **$\bar{X}$-карту** наносят *среднее* каждой выборки, соответствующую центральную линию и контрольные границы, что позволяет наблюдать уровень процесса. На **R-карту** наносят *размах* значений (наибольшее значение минус наименьшее) для каждой выборки, центральную линию и контрольные границы, что позволяет наблюдать изменчивость процесса. Эти и подобные им карты (включая карты процентов) были разработаны В. А. Шехартом (W. A. Shehart).

Почему в карте контроля качества как меру изменчивости иногда используют размах значений вместо (статистически) более удобного стандартного отклонения? Да потому, что размах легче вычислить. Когда компьютеры не были столь широко распространены, это было большим преимуществом, поскольку рабочие самостоятельно вручную могли составлять для себя карты контроля. Кроме того, хотя (как показано в главе 5, "Изменчивость: изучение разнообразия") как мера изменчивости стандартное отклонение лучше, чем размах, для выборки небольшого размера ($n = 4$ или $n = 5$) размах почти так же хорош, как и стандартное отклонение.

Существуют два способа определения контрольных границ и центральной линии для карт каждого из этих двух видов, которые можно использовать в зависимости от наличия некоторых внешних стандартов (например, предыдущего опыта, конструкторской документации или потребительских требований). Если внешний стандарт отсутствует, то границы вычисляют только исходя из данных. При наличии внешнего стандарта границы вычисляют только на основе этого стандарта. В табл. 18.3.1 показан способ вычисления центральной линии и контрольных границ для $\bar{X}$- и R-карт. Множители $A, A_2, d_2, D_1, D_2, D_3, D_4$ приведены

в табл. 18.3.2. Символ $\overline{\overline{X}}$ означает среднее значение всех выборочных средних, а $\overline{R}$ — среднее значение размаха для всех выборок.

Например, предположим, что стандарт отсутствует, а значения выборочных показателей такие, как в табл. 18.3.3.

Таблица 18.3.1. Определение центральной линии и контрольных границ для $\overline{X}$- и R-карт

		Центральная линия	Контрольные границы
$\overline{X}$-карта	Стандарт задан (μ_0 и σ_0)	μ_0	От $\mu_0 - A\sigma_0$ до $\mu_0 + A\sigma_0$
	Стандарт не задан	$\overline{\overline{X}}$	От $\overline{\overline{X}} - A_2\overline{R}$ до $\overline{\overline{X}} + A_2\overline{R}$
R-карта	Стандарт задан (σ_0)	$d_2\sigma_0$	От $D_1\sigma_0$ до $D_2\sigma_0$
	Стандарт не задан	$\overline{R}$	От $D_3\overline{R}$ до $D_4\overline{R}$

Таблица 18.3.2. Множители для построения $\overline{X}$- и R-карт

Размер выборки	Карты для средних ($\overline{X}$-карта)		Карты для диапазонов (R-карта)			
	Множители для контрольных границ		Множитель для центральной линии	Множители для контрольных границ		
n	A	A_2	d_2	D_1	D_2	D_3	D_4
2	2,121	1,880	1,128	0	3,686	0	3,267
3	1,732	1,023	1,693	0	4,358	0	2,574
4	1,500	0,729	2,059	0	4,698	0	2,282
5	1,342	0,577	2,326	0	4,918	0	2,114
6	1,225	0,483	2,534	0	5,078	0	2,004
7	1,134	0,419	2,704	0,204	5,204	0,076	1,924
8	1,061	0,373	2,847	0,388	5,306	0,136	1,864
9	1,000	0,337	2,970	0,547	5,393	0,184	1,816
10	0,949	0,308	3,078	0,687	5,469	0,223	1,777
11	0,905	0,285	3,173	0,811	5,535	0,256	1,744
12	0,866	0,266	3,258	0,922	5,594	0,283	1,717
13	0,832	0,249	3,336	1,025	5,647	0,307	1,693
14	0,802	0,2235	3,407	1,118	5,696	0,328	1,672
15	0,775	0,223	3,472	1,203	5,741	0,347	1,653
16	0,750	0,212	3,532	1,282	5,782	0,363	1,637
17	0,728	0,203	3,588	1,356	5,820	0,378	1,622
18	0,707	0,194	3,640	1,424	5,856	0,391	1,608

Размер выборки	Карты для средних ($\overline{X}$-карта)		Карты для диапазонов (R-карта)			
	Множители для контрольных границ		Множитель для центральной линии	Множители для контрольных границ		
n	A	A_2	d_2	D_1	D_2	D_3	D_4
19	0,688	0,187	3,689	1,487	5,891	0,403	1,597
20	0,671	0,180	3,735	1,549	5,921	0,415	1,585
21	0,655	0,173	3,778	1,605	5,951	0,425	1,575
22	0,640	0,167	3,819	1,659	5,979	0,434	1,566
23	0,626	0,162	3,858	1,710	6,006	0,443	1,557
24	0,612	0,157	3,895	1,759	6,031	0,451	1,548
25	0,600	0,153	3,931	1,806	6,056	0,459	1,541

Значения взяты из ASTM-STP 15D, American Society for Testing and Materials.

Таблица 18.3.3. Значения показателей для восьми выборок из $n = 4$ элементов каждая

Номер выборки	Выборочное среднее, $\overline{X}$	Выборочный размах, R
1	22,3	1,8
2	22,4	1,2
3	21,5	1,1
4	22,0	0,9
5	21,1	1,1
6	21,7	0,9
7	22,1	1,5
8	21,6	1,1
Среднее	$\overline{\overline{X}} = 21,84$	$\overline{R} = 1,20$

Для выборки размером $n = 4$ табличные значения будут следующими: $A_2 = 0,729$; $D_3 = 0$ и $D_4 = 2,282$. Для $\overline{X}$-карты значения центральной линии и предельных границ следующие.

Центральная линия: $\overline{\overline{X}} = 21,84$

Нижняя контрольная граница: $\overline{\overline{X}} - A_2\overline{R} = 21,84 - (0,729)(1,20) = 20,97$

Верхняя контрольная граница: $\overline{\overline{X}} + A_2\overline{R} = 21,84 + (0,729)(1,20) = 22,71$

Для R-карты эти значения следующие.

Центральная линия: $\overline{R} = 1,20$

Нижняя контрольная граница: $D_3\overline{R} = (0)(1,20) = 0$

Верхняя контрольная граница: $D_4\overline{R} = (2,282)(1,20) = 2,74$

Теперь предположим, что в дополнение к данным из табл. 18.3.3 мы имеем стандартное значение (скажем, из прошлого опыта): $\mu_0 = 22{,}0$ и $\sigma_0 = 0{,}50$. Для выборки размером $n = 4$ табличные значения будут следующими: $A = 1{,}500$; $d_2 = 2{,}059$; $D_1 = 0$ и $D_2 = 4{,}698$. Для $\bar{X}$-карты значения центральной линии и предельных границ следующие.

Центральная линия: $\mu_0 = 22{,}00$

Нижняя контрольная граница: $\mu_0 - A\sigma_0 = 22{,}00 - (1{,}500)(0{,}50) = 21{,}25$

Верхняя контрольная граница: $\mu_0 + A\sigma_0 = 22{,}00 + (1{,}500)(0{,}50) = 22{,}75$

Для R-карты эти значения следующие.

Центральная линия: $d_2\sigma_0 = (2{,}059)(0{,}50) = 1{,}03$

Нижняя контрольная граница: $D_1\sigma_0 = (0)(0{,}50) = 0$

Верхняя контрольная граница: $D_2\sigma_0 = (4{,}698)(0{,}50) = 2{,}35$

Пример. *Чистый вес моющего средства для посуды*

Из каждой партии, содержащей 150 пачек моющего средства, выбирают случайным образом 5 пачек и взвешивают их содержимое. Затем исходя из полученных данных составляют карту контроля, которую анализируют, чтобы определить необходимость регулировки процесса. В табл. 18.3.4 приведены значения весов и показатели (среднее, наибольшее, наименьшее и размах) для 25 выборок по 5 пачек в каждой.

Таблица 18.3.4. Вес нетто выборочных пачек моющего средства и выборочные показатели

Номер измерения в выборке	Вес нетто отдельных пачек в каждой выборке, унции					Выборочные показатели			
	1	2	3	4	5	Среднее, $\bar{X}$	Наибольшее	Наименьшее	Размах, R
1	16,12	16,03	16,25	16,19	16,24	16,166	16,35	16,03	0,22
2	16,11	16,10	16,28	16,18	16,16	16,166	16,28	16,10	0,18
3	16,16	16,21	16,10	16,09	16,04	16,120	16,21	16,04	0,17
4	15,97	15,99	16,34	16,18	16,02	16,100	16,34	15,97	0,37
5	16,21	16,00	16,14	16,12	16,10	16,114	16,21	16,00	0,21
6	15,77	16,11	16,01	16,02	16,17	16,016	16,17	15,77	0,40
7	16,02	16,29	16,08	15,96	16,11	16,092	16,29	15,96	0,33
8	15,83	16,08	16,25	16,14	16,15	16,090	16,25	15,83	0,42
9	16,16	15,90	16,08	15,98	16,09	16,042	16,16	15,90	0,26
10	16,08	16,10	16,13	16,03	16,03	16,074	16,13	16,03	0,10
11	15,90	16,16	16,15	15,99	16,07	16,054	16,16	15,90	0,26
12	16,09	16,05	16,07	15,98	15,95	16,028	16,09	15,95	0,14
13	15,98	16,18	16,08	16,08	16,07	16,078	16,18	15,98	0,20

Номер измерения в выборке	Вес нетто отдельных пачек в каждой выборке, унции					Выборочные показатели			
	1	2	3	4	5	Среднее, $\bar{X}$	Наибольшее	Наименьшее	Размах, R
14	16,23	16,05	16,10	16,07	16,16	16,122	16,23	16,05	0,18
15	15,96	16,20	16,35	16,11	16,08	16,140	16,35	15,96	0,39
16	16,00	16,04	16,02	16,03	16,09	16,036	16,09	16,00	0,09
17	16,12	16,12	15,95	15,98	16,10	16,054	16,12	15,95	0,17
18	16,30	16,05	16,10	16,09	16,07	16,122	16,30	16,05	0,25
19	16,11	16,15	16,25	16,03	16,05	16,118	16,25	16,03	0,22
20	15,85	16,06	15,96	16,20	16,25	16,064	16,25	15,85	0,40
21	15,94	15,88	16,02	16,06	16,10	16,000	16,10	15,88	0,22
22	16,15	16,15	16,21	15,95	16,13	16,118	16,21	15,95	0,26
23	16,10	16,17	16,24	16,00	15,87	16,076	16,24	15,87	0,37
24	16,22	16,34	16,40	16,07	16,12	16,230	16,40	16,07	0,33
25	16,32	15,97	15,88	16,03	16,27	16,094	16,32	15,88	0,44
						$\bar{\bar{X}} = 16{,}093$			$\bar{R} = 0{,}263$

Так как часть упаковочного оборудования начали использовать только недавно, то внешние стандарты не известны. Ниже приведены значения центральной линии и контрольных границ для $\bar{X}$- и R-карт, вычисленные с использованием табличных значений для выборки размером $n = 5$.

Для $\bar{X}$-карты.

Центральная линия: $\bar{\bar{X}} = 16{,}093$

Нижняя контрольная граница: $\bar{\bar{X}} - A_2\bar{R} = 16{,}093 - (0{,}577)(0{,}263) = 15{,}941$

Верхняя контрольная граница: $\bar{\bar{X}} + A_2\bar{R} = 16{,}093 + (0{,}577)(0{,}263) = 16{,}245$

Для R-карты.

Центральная линия: $\bar{R} = 0{,}263$

Нижняя контрольная граница: $D_3\bar{R} = (0)(0{,}263) = 0$

Верхняя контрольная граница: $D_4\bar{R} = (2{,}114)(0{,}263) = 0{,}556$

Обе карты (рис. 18.3.1) показывают, что процесс находится под контролем. Все наблюдения на каждой карте находятся в пределах контрольных границ.

Нет ярко выраженных неслучайных закономерностей (таких, например, как направленный вверх или вниз тренд, или слишком много последовательных точек, находящихся слишком близко от границы). Не кажется ли вам подозрительным наблюдение номер 24 в правой части $\bar{X}$, поскольку оно расположено слишком близко к верхней контрольной границе? Возможно, да, но так как оно все же находится внутри

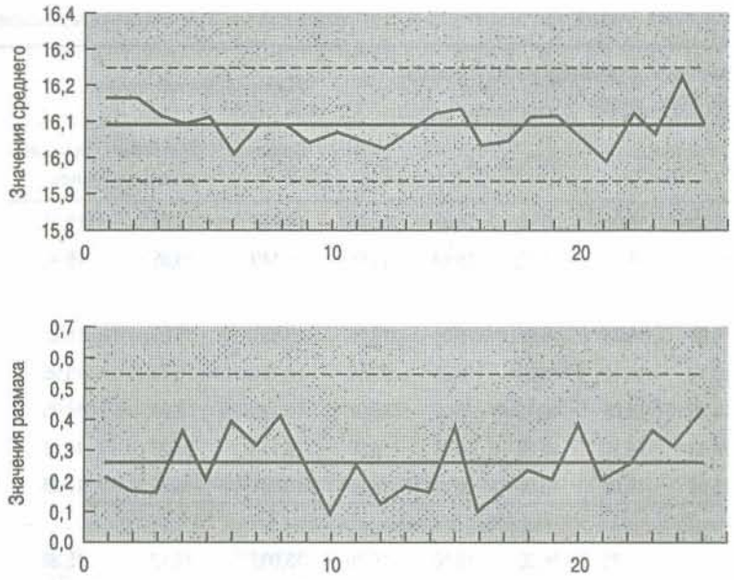
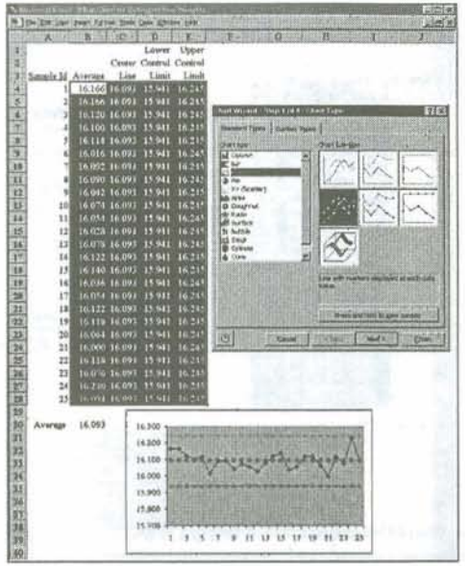


Рис. 18.3.1. $\bar{X}$- и R- карты для 25 выборок, содержащих по пять пачек моющего средства в каждой. Процесс находится под контролем, так как все точки расположены в пределах контрольных границ, а графики демонстрируют случайность изменений без ярко выраженных закономерностей или трендов

границ, лучше не поддаваться искушению и не вмешиваться в процесс упаковки, пока отклонение не станет более явным. Имея только одну точку, расположенную почти на контрольной границе (но в пределах границ), можно продолжать считать, что процесс остается под контролем. Однако следует продолжить наблюдения с целью поиска возможных дальнейших свидетельств и закономерностей, которые покажут необходимость регулировки процесса.

На рис. 18.3.2 показано, как этот метод контроля (с помощью карты контроля) можно было бы применить в цеху. Этот бланк, взятый в Американском обществе контроля качества, позволяет фиксировать общую информацию, результаты измерений, показатели и также содержит карты контроля.

Далее покажем, как использовать Excel для составления $\bar{X}$ -карты веса нетто моющего средства. Начнем с колонки, содержащей список средних значений веса (для выборок из пяти наблюдений). Справа от нее создадим колонку, содержащую многократно повторенное среднее значение $\bar{\bar{X}} = 16,093$. За ней создадим колонку для нижней контрольной границы $\bar{\bar{X}} - A_2\bar{R} = 15,941$ и еще одну — для верхней контрольной границы $\bar{\bar{X}} + A_2\bar{R} = 16,245$. Теперь выделим все эти четыре колонки и, используя команду меню Excel Insert⇨Chart (Вставка⇨Диаграмма) или просто щелкнув на кнопке Chart Wizard (Мастер диаграмм), диалоговое окно мастера диаграмм. В списке Chart type (Тип) выберите Line (График) (с маркерами, помечающими точки данных), как показано ниже на рисунке. Поскольку вы строите график с помощью Мастера диаграмм, щелкните два раза на кнопке Next (Далее) и перейдите к диалоговому окну Chart Options (Параметры диаграммы). Щелкните на вкладке Gridlines tab (Линии сетки), чтобы указать, будете вы использовать координатную сетку или нет (в данном примере она не используется), и затем щелкните на вкладке Legend (Легенда), чтобы удалить пояснения (сняв флажок Show legend (Добавить легенду)). Далее щелкните на кнопке Finish (Готово), и появится $\bar{X}$ -карта.



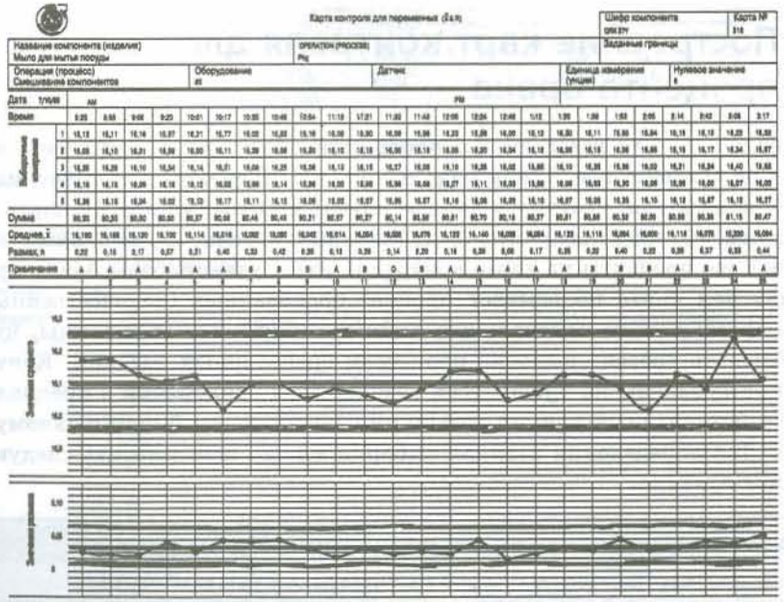
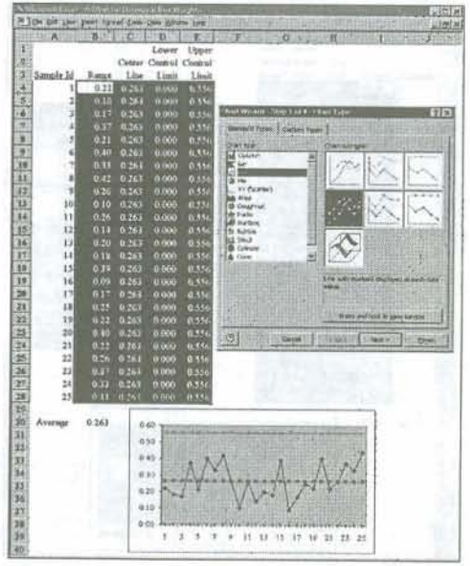


Рис. 18.3.2. Пример фиксации в цеху общей информации, результатов измерений, показателей и карт контроля для веса нетто моющего средства

Чтобы построить с помощью Excel R -карту для данных о моющих средствах, поступают так же, как и при построении $\bar{X}$ -карты, но для первой колонки используют значения размаха R, для второй колонки — их среднее $\bar{R}$, а для третьей и четвертой колонок — нижнюю и верхнюю контрольные границы: $D_3\bar{R} = 0$ и $D_4\bar{R} = 0,556$ соответственно. Ниже показано построение R -карты в Excel.



18.4. Построение карт контроля для процента брака

Предположим, что все единицы продукции проверены и разделены на бракованные и не бракованные. Такая классификация отличается от измерения количественного показателя, поэтому в данном случае необходима другая карта. Для каждой группы единиц продукции подсчитывают процент бракованных единиц и фиксируют эти проценты на карте контроля, чтобы увидеть серьезность проблемы.

Процентная карта показывает процент бракованных (некачественных) изделий, соответствующую центральную линию и контрольные границы, что позволяет наблюдать уровень выпуска процессом бракованных изделий. Контрольные границы установлены на уровне трех стандартных отклонений в предположении, что число бракованных единиц продукции подчиняется биномиальному распределению. Для определения размера выборки можно использовать следующее эмпирическое правило.

Выбор размера выборки, n, для процентной карты контроля

Ожидаемое количество бракованных изделий в выборке должно быть не меньше 5.

Это говорит о том, что для процентной карты необходим больший размер выборки n, чем для $\bar{X}$- и R-карт. Например, если вы ожидаете 10% некачественных изделий, то размер выборки должен быть не менее $n = 5/0{,}10 = 50$. Если вы ожидаете только 0,4% некачественных изделий, то размер выборки должен быть не меньше $n = 5/0{,}004 = 1250$.

Если вы достаточно удачливы и *никогда* не выпускаете некачественные изделия, не отчаивайтесь. Хотя вы не можете определить размер выборки, который

удовлетворяет этому правилу, вы, тем не менее, находитесь в хорошем положении. Примите поздравления по поводу качества вашей продукции!

В табл. 18.4.1 показано, как вычислить центральную линию и контрольные границы для процентной карты. Здесь нет необходимости в специальной таблице множителей, так как эти формулы используют стандартное отклонение биномиального распределения. Напомним, что p обозначает наблюдаемую долю или процент в одной выборке. Символ $\bar{p}$ означает среднее значение всех выборочных значений доли.

Например, предположим, что стандарт не задан, а выборочные значения параметров приведены в табл. 18.4.2.

При размере выборки $n = 500$ центральная линия и контрольные границы вычисляются следующим образом.

Центральная линия:

$$\bar{p} = 0{,}021667 \text{, или } 2{,}1667\%.$$

Таблица 18.4.1. Вычисление центральной линии и контрольных границ для процентной карты

	Центральная линия	Контрольные границы
Стандарт задан (π_0)	π_0	От $\pi_0 - 3\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$ до $\pi_0 + 3\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$
Стандарт не задан	$\bar{p}$	От $\bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$ до $\bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

Таблица 18.4.2. Показатели и результаты измерений для 12 выборок размером $n = 500$ каждая

Номер выборки	Количество единиц брака, X	Выборочный процент, p
1	10	2,0
2	11	2,2
3	10	2,0
4	12	2,4
5	7	1,4
6	14	2,8
7	13	2,6
8	11	2,2
9	6	1,2
10	12	2,4
11	11	2,2
12	13	2,6
Среднее	10,8333	$\bar{p} = 2{,}1667\%$

Нижняя контрольная граница:

$$\bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0,021667 - 3\sqrt{\frac{0,021667(1-0,021667)}{500}} =$$
$$= 0,021667 - (3)(0,006511) = 0,0021,$$

или 0,21%.

Верхняя контрольная граница:

$$\bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0,021667 + (3)(0,006511) = 0,0412, \text{ или } 4,12\%.$$

Теперь предположим, что в дополнение к данным табл. 18.4.2 задано стандартное значение и вам известно (например, из прошлого опыта), что норма брака для этого процесса составляет $\pi_0 = 2,30\%$. Для выборки размером $n = 500$ центральная линия и контрольные границы будут следующими.

Центральная линия:

$$\pi_0 = 0,0230, \text{ или } 2,30\%.$$

Нижняя контрольная граница:

$$\pi_0 - 3\sqrt{\frac{\pi_0(1-\pi_0)}{n}} = 0,0230 - 3\sqrt{\frac{0,0230(1-0,0230)}{500}} =$$
$$= 0,0230 - (3)(0,0067039) = 0,0029,$$

или 0,29%.

Верхняя контрольная граница:

$$\pi_0 + 3\sqrt{\frac{\pi_0(1-\pi_0)}{n}} = 0,0230 + (3)(0,0067039) = 0,0431, \text{ или } 4,31\%.$$

Пример. *Заполнение бланков заказов на покупку товаров*

Когда в компьютер вводят данные о заказах на покупку товаров и обрабатывают их, иногда появляются ошибки, исправление которых требует особого внимания. Конечно, "особое внимание" стоит дорого, и желательно, чтобы процент ошибок был небольшим. Чтобы держать эту проблему под контролем, следует регулярно просматривать процентную карту.

В каждой пачке из 300 заказов фиксируют процент ошибок (табл. 18.4.3).

Поскольку нет никакого стандартного значения для этой процентной карты, центральную линию и контрольные границы вычисляют следующим образом.

Центральная линия:

$$\bar{p} = 0,0515, \text{ или } 5,15\%.$$

Нижняя контрольная граница:

$$\bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0,0515 - 3\sqrt{\frac{0,0515(1-0,0515)}{300}} =$$
$$= 0,0515 - (3)(0,012760) = 0,0132,$$

или 1,32%.

Таблица 18.4.3. Параметры ошибок в 25 пачках, содержащих по 300 заказов в каждой

Номер пачки	Число ошибок, X	Выборочный процент, p
1	5	1,7
2	11	3,7
3	7	2,3
4	14	4,7
5	5	1,7
6	11	3,7
7	11	3,7
8	10	3,3
9	14	4,7
10	8	2,7
11	5	1,7
12	16	5,3
13	12	4,0
14	9	3,0
15	13	4,3
16	17	5,7
17	20	6,7
18	30	10,0
19	23	7,7
20	25	8,3
21	35	11,7
22	24	8,0
23	22	7,3
24	23	7,7
25	16	5,3
Среднее	15,44	$\bar{p} = 5,15\%$

Верхняя контрольная граница:

$$\bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0,0515 + (3)(0,012760) = 0,0898, \text{ или } 8,98\%.$$

Процентная карта (рис. 18.4.1) показывает, что процесс обработки заказов на покупку товаров *не* находится под контролем.

Две пачки (18 и 21) находятся выше верхней контрольной границы, что свидетельствует о значительных ошибках. Кроме того, в правой части карты проценты более высокие, чем в левой. Таким образом, процесс был под контролем с низким уровнем ошибок, но где-то начиная с пачек с номерами 16 и 17 возникла проблема, которая привела к увеличению уровня ошибок.

Карта контроля сделала свое дело: выявила проблему и предоставила данные для ее решения. Да, проблема существует. Об этом свидетельствует высокий уровень ошибок, начиная с пачек с номерами 16 и 17.

Изучение произошедшего при обработке заказов из пачек с номерами 16 и 17 показывает, что в это время началась работа над новым инвестиционным проектом. В связи с этим было так много заказов на товары, что потребовалась нанять дополнительных служащих. Очевидно, повышенный процент ошибок первоначально был обусловлен нагрузкой на систему (из-за большого объема заказов) и недостаточным знанием новых служащих этой системы. Хотя уровень ошибок снижается в конце кривой (правая часть карты), он все еще высокий. В этой ситуации целесообразно провести ускоренный курс подготовки новых служащих, чтобы вернуться к первоначальному невысокому уровню ошибок.

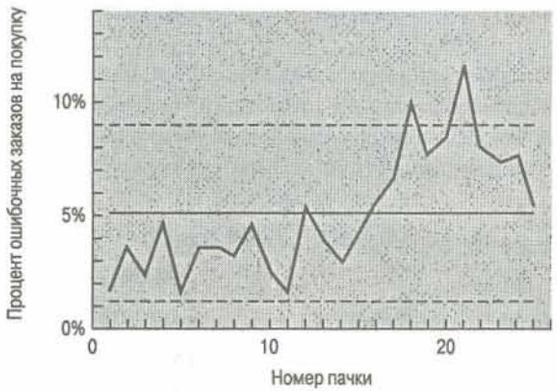


Рис. 18.4.1. Процесс обработки заказов на покупку товаров не находится под контролем и требует внимания. Пачки с номерами 18 и 21 выходят за пределы верхней контрольной границы, а где-то начиная с пачек с номерами 16 и 17 наблюдается сдвиг уровня

18.5. Дополнительный материал

Резюме

Статистический контроль качества — это применение статистических методов для оценки и улучшения результатов любой деятельности. Под **процессом** понимают любой вид экономической деятельности, который превращает ресурсы в продукты. Процесс может состоять из других процессов, называемых *подпроцессами*, каждый из которых также является процессом. **Статистический контроль процесса** заключается в использовании статистических методов для контроля функционирования процесса таким образом, чтобы можно было либо при необходимости регулировать процесс и устранять неполадки, либо не вмешиваться в процесс, если он работает должным образом. Всякий раз, когда есть возможность выяснить, почему возникла та или иная проблема, имеет место **неслучайная причина отклонений**. Все причины отклонений, которые не заслуживают того, чтобы их определять, объединяют под названием **случайные причины отклонений**. Если все неслучайные причины отклонений определены и исключены, так что остались только случайные причины, то говорят, что процесс находится **в состоянии статистического контроля**, или **под контролем**.

Диаграмма Парето показывает причины различных неполадок, которые расположены в порядке от наиболее часто возникающих к наименее часто возникающим, что позволяет сосредоточить внимание на самых важных проблемах. На диаграмме Парето самый высокий столбик будет всегда слева (показывая наиболее часто возникающие проблемы), а самый низкий — справа. Линия кумулятивного процента всегда направлена вверх и при движении вправо имеет тенденцию приближаться к горизонтальной.

Карта контроля содержит последовательность замеров параметров процесса, *центральную линию* и *контрольные границы*, которые вычисляют, чтобы принимать решение о том, находится процесс под контролем или нет. Если принимается решение о том, что процесс не находится под контролем, карта контроля помогает выявить и устранить соответствующую проблему. Проверяемые гипотезы формулируются следующим образом.

- H_0: процесс находится под контролем.
- H_1: процесс не находится под контролем.

Уровень ложной тревоги характеризует, как часто вмешательство в процесс происходит без необходимости. Это понятие аналогично понятию ошибки I рода в проверке гипотез. Общепринятые карты контроля строятся исходя из трех, а не из двух стандартных ошибок параметров, поскольку обычный для статистической проверки гипотез уровень 5% ошибки I рода недопустимо высок для большинства приложений контроля качества. Когда процесс находится под контролем, график на карте контроля не выходит за пределы контрольных границ. Если какое-либо измерение выходит за пределы контрольных границ, либо выше верхней границы, либо ниже нижней границы, то делают вывод, что процесс вышел из-под контроля. Внезапный скачок значений в карте контроля на другой уровень или постепенный сдвиг точек вверх или вниз может также свидетельствовать о том, что процесс не находится под контролем, даже несмотря на то, что все точки расположены в пределах контрольных границ.

Для составления карты количественных измерений обычно выбирают небольшой размер выборки ($n = 4$ или $n = 5$) и затем наносят на карту результаты многих последовательных выборок. **$\bar{X}$-карта** содержит *среднее* каждой выборки, соответствующую центральную линию и контрольные границы, что позволяет контролировать уровень процесса. **R-карта** содержит *размах* (наибольшее значение минус наименьшее) значений для каждой выборки, центральную линию и контрольные границы, что позволяет контролировать изменчивость процесса.

Процентная карта показывает процент бракованных изделий, соответствующую центральную линию и контрольные границы, что позволяет контролировать уровень, на котором процесс начинает производить некачественные изделия. Для построения такой карты размер выборки должен быть больше, чем для карты количественных измерений. Эмпирическое правило состоит в том, что ожидаемое количество бракованных изделий в выборке должно быть не менее пяти.

Основные термины

- Статистический контроль качества (statistical quality control), 908
- Процесс (process), 911
- Статистический контроль процесса (statistical process control), 911
- Неслучайная причина вариации (assignable cause of variation), 912
- Случайная причина отклонения (random causes of variation), 912
- В состоянии статистического контроля (in a state of statistical control), или под контролем (in control), 913
- Диаграмма Парето (Pareto diagram), 913
- Карта контроля (control chart), 915
- Уровень ложной тревоги (false alarm rate), 916
- $\overline{X}$-карта ($\overline{X}$ chart), 919
- R-карта (R chart), 919
- Процентная карта (percentage chart), 926

Контрольные вопросы

1. а) Что такое статистический контроль качества?

   б) Почему статистические методы полезны при контроле качества?
2. а) Что такое процесс?

   б) Какая связь существует между процессом и его подпроцессами?

   в) Что такое статистический контроль процесса?
3. Можно ли применять статистический контроль процесса к бизнес-деятельности в целом или его применение ограничено производством?
4. Для чего осуществляют мониторинг процесса? Не лучше ли просто проверять готовую продукцию и выбраковывать некачественные изделия?
5. а) Что такое неслучайная причина отклонения?

   б) Что такое случайная причина отклонения?
6. а) Что имеют в виду, когда говорят, что процесс находится в состоянии статистического контроля?

   б) Что следует делать, если процесс не контролируется?

   в) Что следует делать, когда выясняется, что процесс находится под контролем?
7. а) Какая информация содержится на диаграмме Парето?

   б) Почему диаграмма Парето является полезным инструментом управления?
8. а) Что такое карта контроля?

   б) Объясните, каким образом карты контроля связаны с выполнением четырех основных функций статистического анализа?

   в) Какие гипотезы проверяют, когда используют карты контроля?

   г) Что такое уровень ложной тревоги? Устанавливают ли его обычно равным 5%?
9. а) Опишите обычную карту контроля для процесса, находящегося под контролем.

   б) Опишите три различных случая, когда на основании карты контроля можно сделать вывод, что процесс не находится под контролем.
10. а) С какой целью применяют $\overline{X}$-карту?

    б) Чему равен обычный размер выборки для такой карты?

    в) Как можно определить центральную линию в случае отсутствия стандарта?

    г) Как можно определить центральную линию, если стандарт известен?

    д) Как можно определить контрольные границы при отсутствии стандарта?

    е) Как можно определить контрольные границы, если стандарт известен?
11. а) С какой целью применяют R-карту?

    б) Чему равен обычный размер выборки для такой карты?

    в) Как можно определить центральную линию при отсутствии стандарта?

    г) Как можно определить центральную линию, если стандарт известен?

    д) Как можно определить контрольные границы при отсутствии стандарта?

    е) Как можно определить контрольные границы, если стандарт известен?
12. а) С какой целью применяют процентную карту?

    б) Насколько большим должен быть размер выборки для процентной карты?

    в) Как можно определить центральную линию при отсутствии стандарта?

    г) Как можно определить центральную линию, если стандарт известен?

    д) Как можно определить контрольные границы при отсутствии стандарта?

    е) Как можно определить контрольные границы, если стандарт известен?

Задачи

1. Укажите, что наилучшим образом подойдет для каждой из приведенных ниже ситуаций — диаграмма Парето, $\overline{X}$-карта, R-карта или процентная карта. Обоснуйте ваш выбор.

   а) Все рабочие хотят вам помочь, но не могут договориться между собой, какие проблемы нужно решать первыми.

   б) Одни двигатели сходят с конвейера со слишком большим количеством масла, а другие — со слишком малым. Необходимо каким-то образом контролировать эту разницу.

   в) Изготовленные механизмы имеют один и тот же размер независимо от партии, но этот размер имеет тенденцию быть больше того, который требуется в соответствии со спецификацией.

   г) Руководство хотело бы определить уровень производства некачественных оберток для конфет.

д) Вам бы хотелось узнать пределы колебаний во время работы параметров оборудования с целью отрегулировать его таким образом, чтобы каждая бутылка заполнялась чуть больше, чем указано на этикетке.

е) Обычно вы оплачиваете счета вовремя, но небольшая их часть проходит через систему с ошибкой и оплачивается с запозданием, что приводит к штрафу. Вы хотите держать этот процесс под контролем, чтобы знать, когда дела ухудшаются.

2. У тракторного завода проблемы с отделением, которое изготавливает трансмиссии. На рис. 18.5.1 показана диаграмма Парето, построенная по последним данным.

а) Какая проблема является наиболее важной с точки зрения влияния на качество трансмиссий? Какой процент всех трудностей связан с этой проблемой?

б) Какая проблема является следующей по значимости? Какой процент всех трудностей связан с этой проблемой?

в) Какой суммарный процент некачественных трансмиссий связан с двумя наиболее важными проблемами?

г) Какой суммарный процент некачественных трансмиссий связан с тремя наиболее важными проблемами?

д) Напишите своему инспектору служебную записку с обобщением данных об этой ситуации и рекомендациями необходимых действий.

3. Производитель кондитерских изделий отслеживает частоту появления различных проблем при производстве небольших шоколадок с твердым карамельным покрытием. Основные данные представлены в табл. 18.5.1.

а) Расположите проблемы в порядке уменьшения частоты их появления (от наиболее к наименее часто появляющейся) и составьте таблицу, показывающую количество появлений, процент от общего количества случаев и кумулятивный процент.

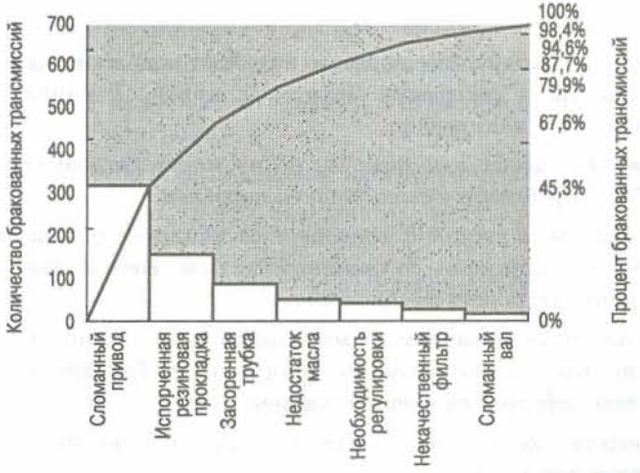


Рис. 18.5.1. Диаграмма Парето для некачественных трансмиссий

Таблица 18.5.1. Частота появления различных проблем на кондитерской фабрике

Причина проблемы	Количество случаев
Неоднородность плитки	22
Недостаточная толщина покрытия	526
Раздавленная плитка	292
Слишком большая толщина покрытия	89
Две слипшиеся вместе плитки	57

б) Постройте диаграмму Парето для этой ситуации.

в) Какая проблема является наиболее важной с точки зрения количества некачественных плиток шоколада? Какой процент всех трудностей связан с этой проблемой?

г) Какая проблема является следующей по значимости? Какой процент всех трудностей связан с этой проблемой?

д) Какой суммарный процент бракованных плиток шоколада связан с двумя наиболее важными проблемами?

е) Напишите своему инспектору служебную записку с обобщением данных об этой ситуации и рекомендациями необходимых действий.

4. Фирма, специализирующаяся на оформлении сертификатов о скидке цены, свела в виде таблицы частоты появления различных проблем (табл. 18.5.2).

а) Расположите проблемы от наиболее к наименее часто встречающейся и составьте таблицу, содержащую количество случаев возникновения проблемы, процент от общего количества проблем и кумулятивный процент.

б) Постройте диаграмму Парето для этой ситуации.

в) Какая проблема является самой важной с точки зрения влияния на сертификаты? Какой процент всех трудностей связан с этой проблемой?

г) Какая проблема является следующей по значимости? Какой процент всех трудностей связан с этой проблемой?

д) Какой суммарный процент неверно оформленных сертификатов связан с двумя наиболее важными проблемами?

е) Напишите своему инспектору служебную записку с обобщением данных об этой ситуации и рекомендациями необходимых действий.

Таблица 18.5.2. Частота появления различных проблем в оформлении сертификатов

Причина проблемы	Количество случаев
Незаполненный	53
Неразборчивый текст	528
Два числа, переставленных местами	184
Неверное место в бланке	330

5. Рассмотрите $\bar{X}$-карту, приведенную на рис. 18.5.2, которая содержит среднее количество шоколадных стружек на одно печенье.

   а) Опишите в общих чертах, что вы видите на этой карте.

   б) Определите, находится ли процесс под контролем? Обоснуйте ваш ответ.

   в) Какое действие, если такое есть, является обязательным?

6. Определите центральную линию и контрольные границы для каждой из следующих ситуаций.

   а) $\bar{X}$-карта, размер выборки $n = 6$, $\bar{\bar{X}} = 56{,}31$; $\bar{R} = 4{,}16$; стандарт не задан.

   б) R-карта, размер выборки $n = 6$, $\bar{\bar{X}} = 56{,}31$; $\bar{R} = 4{,}16$; стандарт не задан.

   в) $\bar{X}$-карта, размер выборки $n = 3$, $\bar{\bar{X}} = 182{,}3$; $\bar{R} = 29{,}4$; стандарт не задан.

   г) R-карта, размер выборки $n = 3$, $\bar{\bar{X}} = 182{,}3$; $\bar{R} = 29{,}4$; стандарт не задан.

   д) $\bar{X}$-карта, размер выборки $n = 5$, $\bar{\bar{X}} = 108{,}3$; $\bar{R} = 13{,}8$; заданы следующие значения стандартов: $\mu_0 = 100{,}0$ и $\sigma_0 = 5{,}0$.

   е) R-карта, размер выборки $n = 5$, $\bar{\bar{X}} = 108{,}3$; $\bar{R} = 13{,}8$; заданы следующие значения стандартов: $\mu_0 = 100{,}0$ и $\sigma_0 = 5{,}0$.

   ж) $\bar{X}$-карта, размер выборки $n = 8$, заданы следующие значения стандартов: $\mu_0 = 2{,}500$ и $\sigma_0 = 0{,}010$.

   з) R-карта, размер выборки $n = 8$, заданы следующие значения стандартов: $\mu_0 = 2{,}500$ и $\sigma_0 = 0{,}010$.

7. Рассмотрите представленные в табл. 18.5.3 данные о толщине защитного покрытия.

   а) Вычислите среднее значение $\bar{X}$ и размах R для каждой выборки.

   б) Вычислите общее среднее $\bar{\bar{X}}$ и среднее значение размаха $\bar{R}$.

   в) Определите центральную линию для $\bar{X}$-карты.

   г) Определите контрольные границы для $\bar{X}$-карты.

   д) Постройте $\bar{X}$-карту.

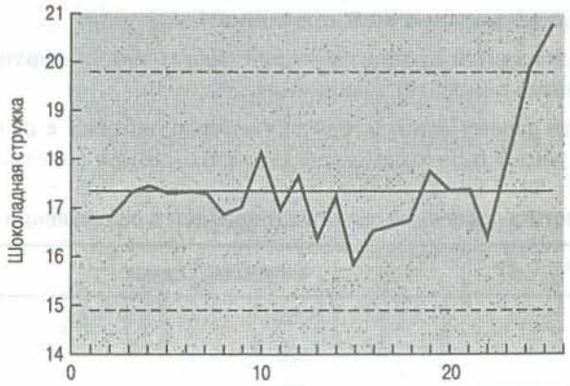


Рис. 18.5.2. $\bar{X}$-карта, показывающая количество шоколадных стружек на одно печенье

Таблица 18.5.3. Толщина защитного покрытия: 25 выборок по 3 единицы в каждой

Номер в выборке	Результаты измерений для отдельных единиц в каждой выборке		
	1	2	3
1	12,51	12,70	12,57
2	12,60	12,53	12,39
3	12,40	12,81	12,56
4	12,44	12,57	12,60
5	12,78	12,61	12,58
6	12,75	12,43	12,61
7	12,53	12,51	12,68
8	12,64	12,49	12,51
9	12,57	12,74	12,81
10	12,70	12,87	12,95
11	12,74	12,80	12,86
12	12,90	12,83	12,91
13	13,05	13,00	13,00
14	12,88	12,88	13,11
15	13,03	12,85	13,05
16	12,96	12,88	12,95
17	12,91	12,75	13,01
18	12,95	13,03	12,89
19	13,17	12,81	13,17
20	13,17	13,05	12,97
21	12,95	13,04	12,80
22	13,04	13,25	12,95
23	13,12	13,07	13,11
24	12,83	13,13	13,31
25	13,24	13,18	13,13

   е) Прокомментируйте $\bar{X}$-карту. В частности, определите, находится ли процесс под контролем? Как вы это узнали?

   ж) Напишите своему инспектору служебную записку с обобщением данных об этой ситуации и рекомендациями необходимых действий.

8. Продолжите работу с данными о толщине защитного покрытия в табл. 18.5.3.

   а) Определите центральную линию для R-карты.

   б) Определите контрольные границы для R-карты.

в) Постройте R-карту.

г) Прокомментируйте R-карту. В частности, определите, находится ли изменчивость процесса под контролем? Как вы это узнали?

9. Мистера К. Р. Вуда, президента Broccoli Enterprises, интересуют приведенные *в табл. 18.5.4 данные о длине стеблей брокколи после обрезки*.

а) Вычислите среднее значение $\bar{X}$ и размах R для каждой выборки.

б) Вычислите общее среднее $\bar{\bar{X}}$ и среднее значение размаха $\bar{R}$.

в) Определите центральную линию для $\bar{X}$-карты.

г) Определите контрольные границы для $\bar{X}$-карты.

д) Постройте $\bar{X}$-карту.

е) Прокомментируйте $\bar{X}$-карту. В частности, определите, находится ли процесс под контролем? Как вы это узнали?

ж) Напишите краткий отчет с обобщением данных об этой ситуации и обоснованием необходимых с вашей точки зрения действий.

Таблица 18.5.4. Длина стеблей брокколи: 20 выборок по 4 стебля в каждой

Номер в выборке	Результаты измерения длины отдельных стеблей в каждой выборке			
	1	2	3	4
1	8,60	8,47	8,44	8,51
2	8,43	8,42	8,62	8,46
3	8,65	8,32	8,65	8,51
4	8,39	8,54	8,50	8,41
5	8,49	8,53	8,61	8,46
6	8,63	8,46	8,64	8,54
7	8,47	8,63	8,54	8,55
8	8,52	8,50	8,31	8,63
9	8,35	8,43	8,51	8,61
10	8,31	8,65	8,46	8,40
11	8,58	8,43	8,55	8,45
12	8,28	8,57	8,58	8,48
13	8,45	8,52	8,52	8,54
14	8,38	8,48	8,41	8,57
15	8,56	8,60	8,58	8,51
16	8,39	8,47	8,59	8,41
17	8,53	8,58	8,54	8,42
18	8,78	8,52	8,46	8,50
19	8,48	8,49	8,74	8,59
20	8,46	8,47	8,70	8,32

10. Продолжите работу с данными о длине стеблей брокколи после обрезки (табл. 18.5.4).

а) Определите центральную линию для R-карты.

б) Определите контрольные границы для R-карты.

в) Постройте R-карту.

г) Прокомментируйте R-карту. В частности, определите, находится ли изменчивость процесса под контролем? Как вы это узнали?

11. Определите центральную линию и значения контрольных границ для каждой из следующих ситуаций.

а) Процентная карта, размер выборки $n = 300$, $\bar{p} = 0{,}0731$.

б) Процентная карта, размер выборки $n = 450$, $\bar{p} = 0{,}1683$.

в) Процентная карта, размер выборки $n = 800$, $\bar{p} = 0{,}0316$, задано значение стандарта $\pi_0 = 0{,}0350$.

г) Процентная карта, размер выборки $n = 1500$, задано значение стандарта $\pi_0 = 0{,}01$.

12. Рассмотрите представленные в табл. 18.5.5 данные о количестве ошибок в пачках из 500 счетов на оплату.

а) Вычислите для каждой пачки процент ошибок p.

б) Вычислите среднее значение процента $\bar{p}$.

в) Найдите центральную линию для процентной карты.

г) Вычислите контрольные границы для процентной карты.

д) Постройте процентную карту.

Таблица 18.5.5. *Количество ошибочных счетов*: 25 пачек размером N= 500

Номер пачки	Количество ошибок, X	Номер пачки	Количество ошибок, X
1	58	14	51
2	57	15	54
3	60	16	47
4	64	17	52
5	57	18	50
6	53	19	62
7	53	20	56
8	74	21	60
9	40	22	67
10	54	23	50
11	56	24	60
12	54	25	67
13	60		

е) Прокомментируйте процентную карту. В частности, определите, находится ли процесс под контролем? Как вы это узнали?

ж) Напишите своему инспектору служебную записку с обобщением данных об этой ситуации и рекомендациями необходимых действий.

13. Независимо от того, насколько строго контролируется производственный процесс, некоторые микросхемы будут работать быстрее, а значит, и стоить дороже. Цель состоит в том, чтобы таких микросхем было как можно больше, поэтому процесс постоянно совершенствуется. Рассмотрите представленные в табл. 18.5.6 данные о количестве микросхем более высокой производительности. Приведены данные о 25 партиях по 1000 микросхем в каждой.

Таблица 18.5.6. Число высокоскоростных микросхем памяти в 25 партиях по 1000 микросхем в каждой

Номер партии	Число высокоскоростных микросхем, X	Доля высокоскоростных микросхем в партии, p
1	75	0,075
2	61	0,061
3	62	0,062
4	70	0,070
5	60	0,060
6	56	0,056
7	61	0,061
8	65	0,065
9	54	0,054
10	71	0,071
11	84	0,084
12	84	0,084
13	110	0,110
14	71	0,071
15	103	0,103
16	103	0,103
17	80	0,080
18	90	0,090
19	84	0,084
20	88	0,088
21	111	0,111
22	118	0,118
23	147	0,147
24	136	0,136
25	123	0,123
Среднее	86,68	0,08668

а) Найдите центральную линию для процентной карты.

б) Вычислите контрольные границы для процентной карты.

в) Постройте процентную карту.

г) Прокомментируйте процентную карту. В частности, определите, находится ли процесс под контролем? Как вы это узнали?

д) Напишите своему инспектору служебную записку с обобщением данных об этой ситуации и рекомендациями необходимых действий.

14. Рассмотрите представленные в табл. 18.5.7 результаты измерения температуры в печи, выполненные четыре раза в час.

а) Постройте для каждого дня $\bar{X}$- и R-карты.

Таблица 18.5.7. Среднее и размах значений температуры, измеряемой 4 раза в час

Время	Понедельник		Вторник		Среда	
	$\bar{X}$	R	$\bar{X}$	R	$\bar{X}$	R
12:00	408,65	30,74	401,07	25,23	402,92	31,96
1:00	401,57	24,81	405,97	32,72	407,28	9,11
2:00	395,52	21,93	401,70	34,56	399,61	22,85
3:00	402,25	35,91	402,06	38,15	398,43	38,52
4:00	405,04	28,68	403,35	31,03	389,97	12,16
5:00	404,12	38,18	407,82	34,93	402,37	18,39
6:00	404,44	18,16	400,30	30,56	406,29	48,44
7:00	407,19	14,14	403,69	17,97	407,77	32,63
8:00	407,43	21,56	399,72	14,11	398,22	19,30
9:00	412,60	25,29	394,77	28,89	408,42	19,11
10:00	413,40	14,32	400,82	37,26	402,91	28,52
11:00	407,26	39,70	401,96	33,30	391,20	20,08
12:00	402,97	32,92	399,94	16,43	398,59	13,29
1:00	387,44	21,89	401,01	16,95	401,72	35,90
2:00	414,39	14,34	399,67	30,34	394,37	12,32
3:00	401,25	18,62	401,67	29,53	409,59	32,91
4:00	400,43	27,96	413,30	12,62	421,97	40,98
5:00	399,31	25,93	412,47	45,47	394,58	48,70
6:00	403,14	37,57	406,62	43,65	407,01	25,75
7:00	403,07	33,52	421,90	21,75	403,40	63,81
8:00	403,66	45,69	429,67	24,30	404,93	82,12
9:00	404,05	40,52	422,75	25,79	391,82	67,03
10:00	399,00	39,77	422,56	15,28	393,96	84,53
11:00	410,18	37,71	424,39	16,64	421,68	92,92

б) Проанализируйте эти карты. В частности, определите, находился ли процесс в этот день под контролем? Как вы это узнали?

в) Для каждого дня укажите, какие, по вашему мнению, мероприятия целесообразно провести?

г) Для изготовления нового продукта необходима постоянная температура с отклонением плюс-минус 10 градусов. На основании контрольных карт из п. "а" скажите, можно ли для изготовления такого продукта использовать эту печь? Обоснуйте свой ответ.

15. Для процесса, который находится под контролем, найдите вероятность того, что конкретный набор из восьми последовательных точек попадет по одну из сторон центральной линии. (*Подсказка*: для процесса, который находится под контролем, считать, что для точки вероятность оказаться по ту же сторону, что и первая точка, равна 0,5 и все точки независимы. Следовательно, можно вычислить вероятность для биномиального распределения при $\pi = 0{,}5$ и $n = 7$. Следует использовать $n = 7$, а не $n = 8$, поскольку первая точка в этой последовательности может оказаться по любую сторону от центральной линии, т.е. ситуацию фактически определяют остальные семь точек.)

16. Какие проблемы (если они есть) видны на картах контроля, приведенных на рис. 18.5.3? Какие мероприятия (если необходимо) вы можете предложить?

17. Какие проблемы (если они есть) видны на картах контроля, приведенных на рис. 18.5.4? Какие мероприятия (если необходимо) вы можете предложить?

18. Какие проблемы (если они есть) видны на картах контроля, приведенных на рис. 18.5.5? Какие мероприятия (если необходимо) вы можете предложить?

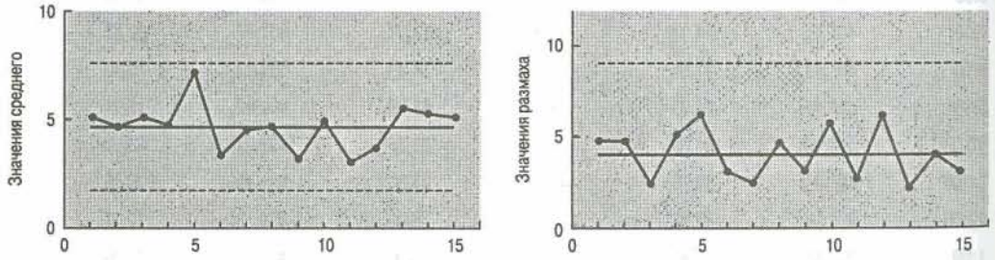


Рис. 18.5.3

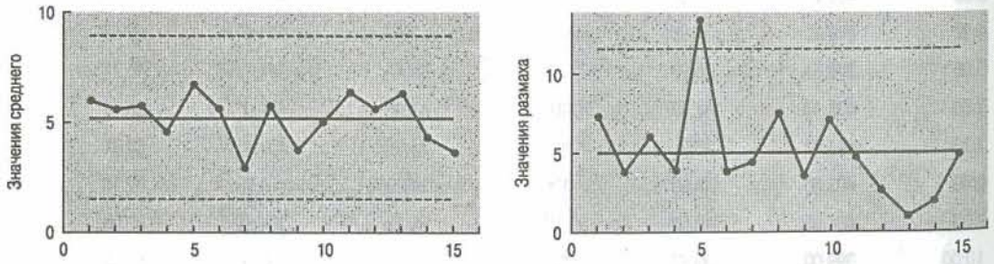


Рис. 18.5.4

19. Какие проблемы (если они есть) видны на картах контроля, приведенных на рис. 18.5.6? Какие мероприятия (если необходимо) вы можете предложить?

20. Какие проблемы (если они есть) видны на картах контроля, приведенных на рис. 18.5.7? Какие мероприятия (если необходимо) вы можете предложить?

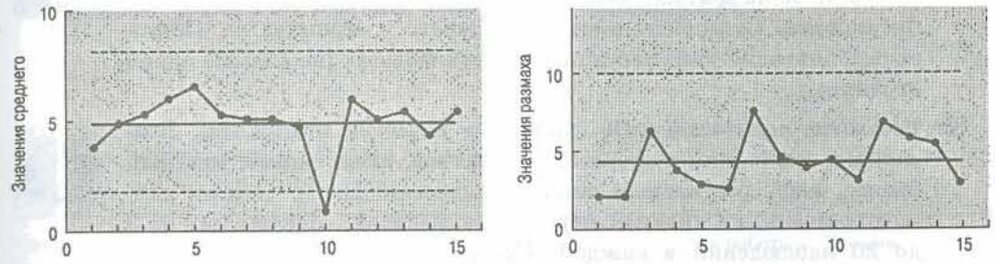


Рис. 18.5.5

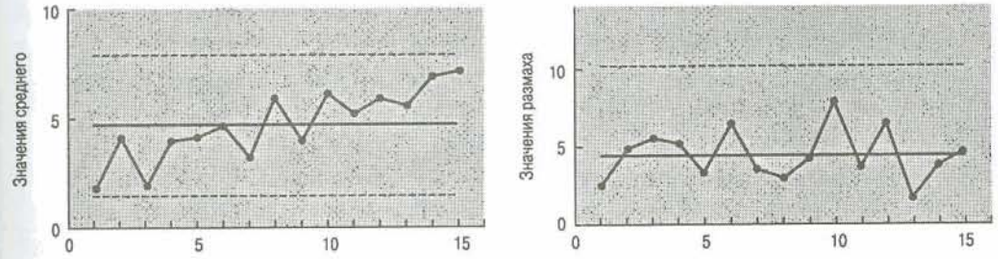


Рис. 18.5.6

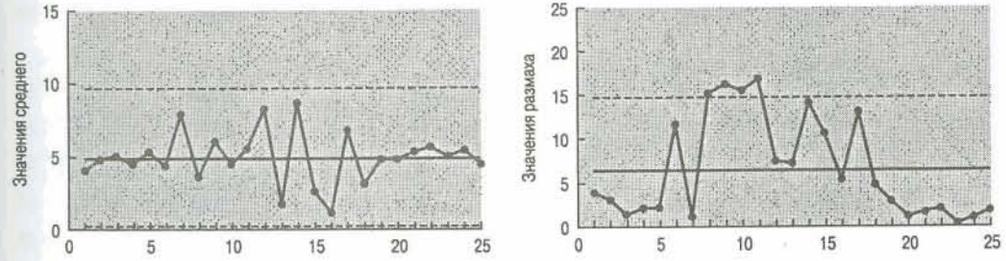


Рис. 18.5.7

Проекты



1. Возьмите качественные данные о том, как часто возникают определенные ситуации. Возможные источники информации: Internet, ваша фирма, местный бизнес, ваш собственный опыт, библиотека. Постройте диаграмму Парето и опишите ее. Подготовьте описание этой ситуации для руководства (не больше одной страницы).



2. Возьмите количественные данные о качестве продукции. Возможные источники информации: Internet, ваша фирма, местный бизнес, ваш собственный опыт, библиотека. Набор данных должен состоять, по крайней мере, из 10 выборок, содержащих от 3 до 20 наблюдений в каждой. Постройте и опишите $\bar{X}$- и R-карты. Подготовьте одну страницу с описанием этой ситуации для руководства.

База данных служащих

Ниже представлены данные о служащих административного отдела по состоянию на 6 января 1999 года.

Номер служащего[1]	Заработная плата за год, дол.	Пол	Возраст, лет	Стаж работы, г.	Уровень подготовки[2]
1	32368	Ж	42	3	B
2	53174	M	54	10	B
3	52722	M	47	10	A
4	53423	M	47	1	B
5	50602	M	44	5	B
6	49033	M	42	10	A
7	24395	M	30	5	A
8	24395	Ж	52	6	A
9	43124	M	48	8	A
10	23975	Ж	58	4	A
11	53174	M	46	4	C
12	58515	M	36	8	C
13	56194	M	49	10	B
14	49033	Ж	55	10	B
15	44884	M	41	1	A
16	53429	Ж	52	5	B
17	46574	M	57	8	A
18	58968	Ж	61	10	B
19	53174	M	50	5	A
20	53627	M	47	10	B

Продолжение таблицы

Номер служащего[1]	Заработная плата за год, дол.	Пол	Возраст, лет	Стаж работы, г.	Уровень подготовки[2]
21	49033	М	54	5	B
22	54981	М	47	7	A
23	62530	М	50	10	B
24	27525	Ж	38	3	A
25	24395	М	31	5	A
26	56884	М	47	10	A
27	52111	М	56	5	A
28	44183	Ж	38	5	B
29	24967	Ж	55	6	A
30	35423	Ж	47	4	A
31	41188	Ж	35	2	B
32	27525	Ж	35	3	A
33	35018	М	39	1	A
34	44183	М	41	2	A
35	35423	М	44	1	A
36	49033	М	53	8	A
37	40741	М	47	2	A
38	49033	М	42	10	A
39	56294	Ж	44	6	C
40	47180	Ж	45	5	C
41	46574	М	56	8	A
42	52722	М	38	8	C
43	51237	М	58	2	B
44	53627	М	52	8	A
45	53174	М	54	10	A
46	56294	М	49	10	B
47	49033	Ж	53	10	B
48	49033	М	43	9	A
49	55549	М	35	8	C

Окончание таблицы

Номер служащего[1]	Заработная плата за год, дол.	Пол	Возраст, лет	Стаж работы, г.	Уровень подготовки[2]
50	51237	М	56	1	C
51	35200	Ж	38	1	B
52	50174	Ж	42	5	A
53	24352	Ж	35	1	A
54	27525	Ж	40	3	A
55	29606	Ж	34	4	B
56	24352	Ж	35	1	A
57	47180	Ж	45	5	B
58	49033	М	54	10	A
59	53174	М	47	10	A
60	53429	Ж	45	7	B
61	53627	М	47	10	A
62	26491	Ж	46	7	A
63	42961	М	36	3	B
64	53174	М	45	5	A
65	36292	М	46	0	A
66	37292	М	47	1	A
67	41188	Ж	34	3	B
68	57242	Ж	45	7	C
69	53429	Ж	44	6	C
70	53174	М	50	10	B
71	44138	Ж	38	2	B

[1] Эти числа указаны с единственной целью – присвоить каждому служащему свой уникальный номер.

[2] Периодически и в добровольном порядке служащему предлагают пройти курс подготовки (это не является обязательным требованием). Служащие, не прошедшие подготовку, получают квалификацию "A", после прохождения одного курса служащий получает квалификацию "B", после второго и заключительного курса – квалификацию "C".

Самопроверка: решение некоторых задач, а также упражнений, использующих базу данных

Глава 1

Задача

6. а) Предварительное исследование данных. Данные уже есть (они изучались раньше), поэтому мы имеем дело не с планированием исследования. Нет также никаких указаний на то, что понадобится оценивать или проверять гипотезы.

   б) Планирование исследования. Поскольку данных еще нет, то нельзя вести речь ни о каких других этапах анализа.

   а) Проверка гипотезы. Есть две возможности: либо разница в заработной плате объясняется только простой случайностью, либо это не так.

   г) Оценивание. Значение валового национального продукта в следующем квартале представляет собой неизвестную количественную величину, относительно которой, исходя из имеющихся данных, нужно сделать предположение.

Глава 2

Задача

9. а) Отдельный служащий представляет собой элементарную единицу этого набора данных.

   б) Это многомерный набор данных, содержащий не менее трех колонок.

   в) Количественные переменные — размер заработной платы за год и стаж работы, качественные переменные — пол и образование.

   г) Образование является порядковой качественной переменной, поскольку естественное упорядочение значений HS, BA, MBA соответствует возрастанию уровня образования.

   д) Это данные об одном временном срезе, в них отсутствует какая-либо последовательность.

Упражнения с использованием базы данных

1. а) Этот набор данных многомерный, поскольку содержит не меньше трех колонок.

   б) Уровень подготовки представляет собой порядковую переменную, поскольку значения этой переменной можно упорядочить содержательным образом.

2. Для пола:

   а) Нет, эти категории нельзя складывать или вычитать в том виде, в каком они представлены в базе данных.

   б) Да, можно подсчитывать количество мужчин или женщин.

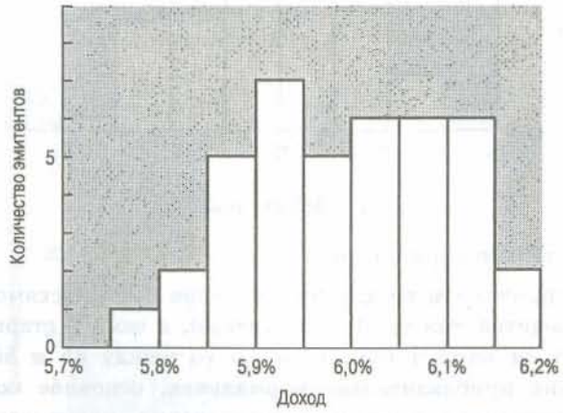
   в) Нет, этот порядок нельзя рассматривать как естественный (содержательно обоснованный).

   г) Да, можно вычислять процент мужчин или женщин.

Глава 3

Задачи

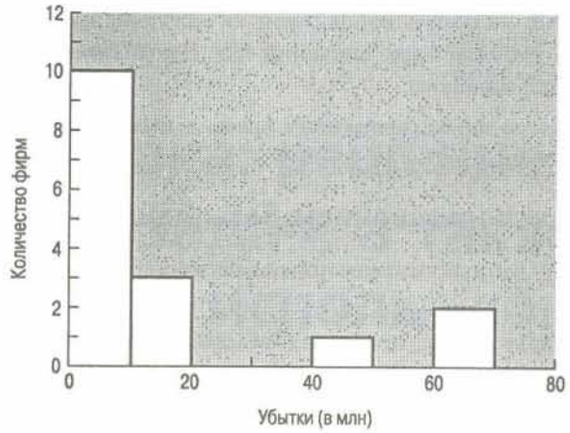
6. а)



   б) Типические значения находятся приблизительно между 5,8 и 6,2%.

   в) Приблизительно нормальное.

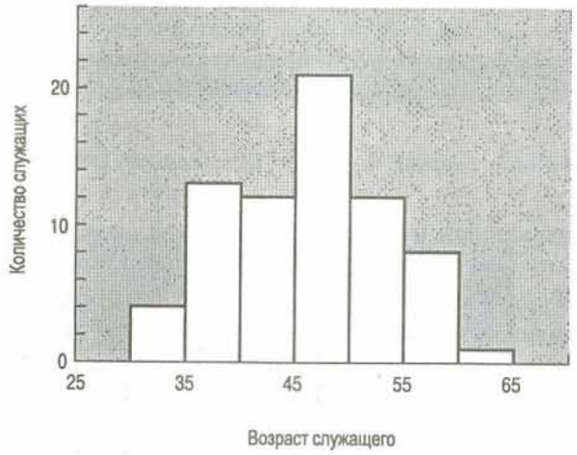
17. а)



б) Распределение скошено в сторону более высоких значений и демонстрирует два разрыва и три выброса. В частности, 13 значений из 16 находятся в первых двух столбиках.

Упражнения с использованием базы данных

2. а)



б) Приблизительно нормальное.

в) Из гистограммы мы видим, что значение возраста самого молодого служащего находится между 30 и 35 годами, а самого старшего между 60 и 65. Типический возраст находится где-то между 45 и 50 годами. Форма распределения приблизительно нормальная, основное количество служащих концентрируется в центре распределения, количество молодых и пожилых служащих относительно невелико.

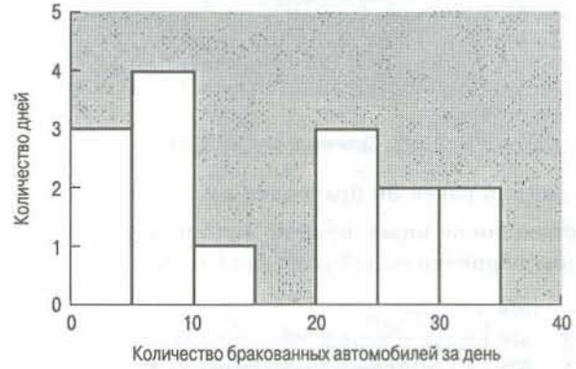
Глава 4

Задачи

1. а) Среднее равно 15,6 бракованных автомобилей в день.

б) Значение медианы равно 14 бракованным автомобилям в день.

в)

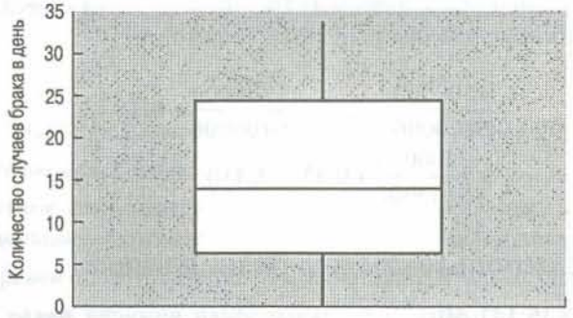


г) Мода равна 7,5 бракованных автомобилей в день. (Для количественных данных значение моды определяют как значение наивысшей точки на гистограмме, или, как правило, как среднюю точку самого высокого столбца гистограммы.) Для другой гистограммы (например, гистограммы со столбцами различной ширины) значение моды может быть другим.

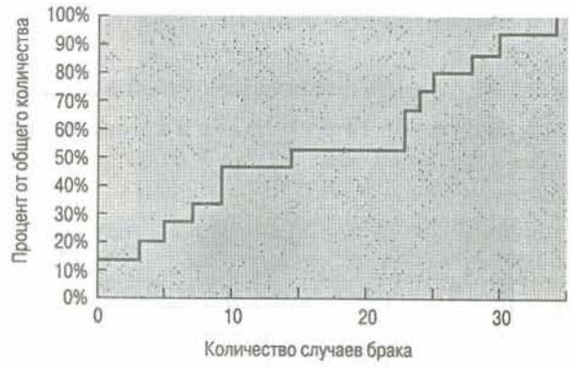
д) Значение нижнего квартиля равно 16 бракованным автомобилям в день, верхнего — 24,5.

е) Наименьшее значение равно 0, наибольше — 34 бракованным автомобилям в день.

ж)

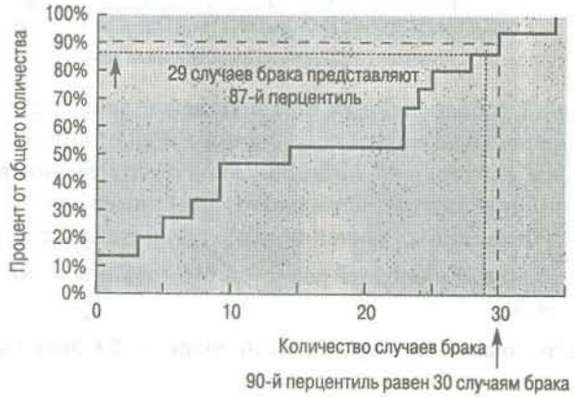


з)



и) 90-й перцентиль равен 30 бракованным автомобилям в день.

к) Из представленной ниже кривой функции кумулятивного распределения видно, что перцентильный ранг приблизительно равен 87%.



6. Стоимость капитала составляет 14,6% и вычисляется как взвешенное среднее значение ставок дохода (17%, 13% и 11%) с весами, равными соответствующим значениям рыночной стоимости:

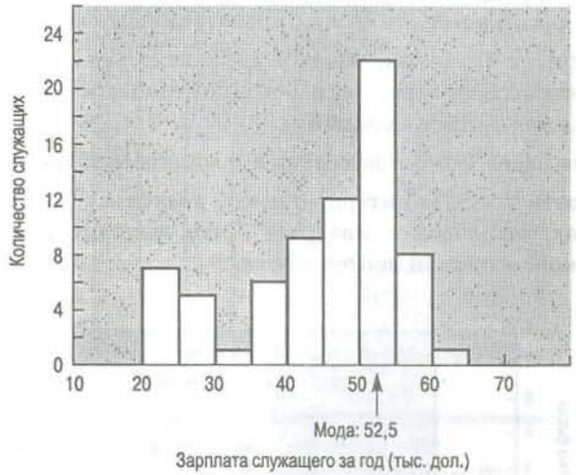
$$\frac{4500000}{8400000}\times 0{,}17+\frac{1700000}{8400000}\times 0{,}13+$$
$$+\frac{2200000}{8400000}\times 0{,}11=0{,}416$$

Упражнения с использованием базы данных

1. а) Среднее: $45 141,50.

   б) Медиана: $49 033.

в)



Мода: приблизительно $52 500, середина наивысшего столбца этой гистограммы.

г) Самое низкое значение имеет среднее, далее следует медиана, наибольшее значение имеет мода. Такое соотношение следует ожидать в случае асимметричного распределения, затянутого в сторону низких значений.

Среднее, рассчитанное как частное от деления суммы всех размеров заработной платы на общее число служащих, равно $45 141. Если каждый служащий получит среднюю заработную плату, то общий объем денег для выплаты не изменится. Медиана свидетельствует о том, что количество служащих, получающих больше $49 030 в год, равно количеству тех, кто получают меньше $49 030 в год. Мода показывает, что в этой фирме наибольшее количество служащих получают от $50 000 до $55 000 в год, если сравнивать с любым другим интервалом заработной платы шириной $5000.

Глава 5

Задачи

1. а) Среднее значение размера бюджета равно $49 500 000.

   б) Стандартное отклонение равно $55 750 000.

   в) Стандартное отклонение приблизительно показывает, насколько сильно размер отдельного бюджета отклоняется от среднего значения.

   г) Размах равен $185 000 000, что вычисляется как (200 –15).

   д) Размах равен разнице наибольшего и наименьшего значений. Фирма с наибольшим бюджетом тратит на $185 000 000 больше, чем фирма с наименьшим бюджетом.

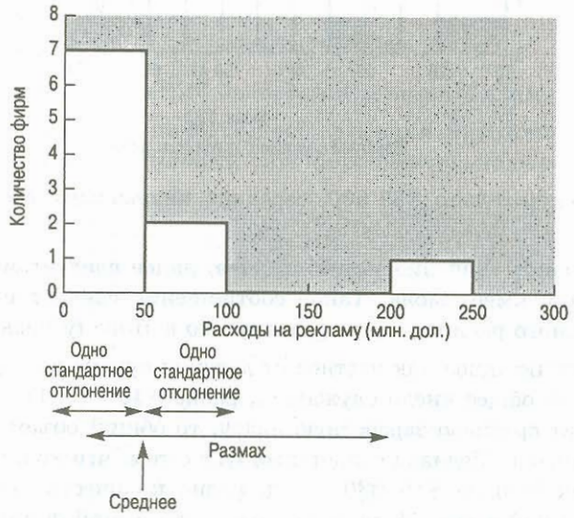
е) Коэффициент вариации равен 1,13 и вычисляется как 55,75/49,50. Коэффициент вариации является безразмерной величиной (просто число) и не зависит от того, в каких единицах производились вычисления.

ж) Коэффициент вариации, равный 1,13, показывает, что бюджет на проведение рекламы для этих фирм обычно отличается от среднего значения на 113% (т.е. на 113% от среднего).

з) Дисперсия равна 3,108 и измеряется в миллионах "долларов в квадрате".

и) Нельзя дать простую интерпретацию, поскольку дисперсия выражается в миллионах "долларов в квадрате", что выходит за пределы обычной практики экономической деятельности.

к)



6. а) Среднее количество административных сотрудников в одной фирме равно 10,4.

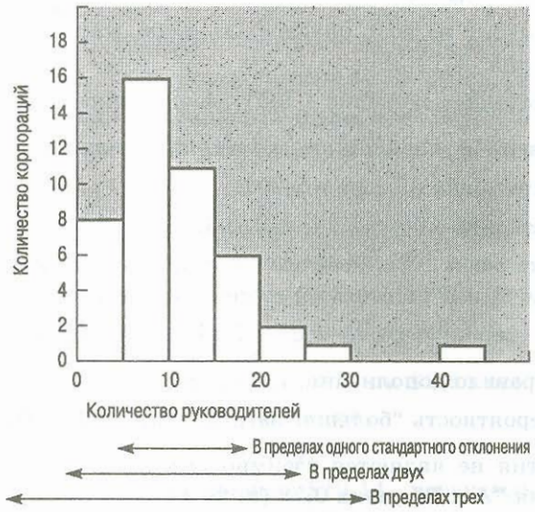
б) Стандартное отклонение, равное 7,19, показывает, что в этих фирмах количество административных сотрудников отличается от среднего значения приблизительно на 7,19 человек.

в) В пределах одного стандартного отклонения от среднего (т.е. в интервале от 3,21 до 17,59) находится 38 корпораций (84,4%). Это больше, чем две трети, что можно было ожидать в случае нормального распределения.

г) В пределах двух стандартных отклонений от среднего (т.е. в интервале от −3,21 до 24,77) находится 43 корпорации (95,6%). Это достаточно близко к значению 95%, которое можно было бы ожидать для нормального распределения.

д) В пределах трех стандартных отклонений от среднего (т.е. в интервале от −11,16 до 31,96) находится 44 корпорации (97,8%). Это достаточно близко к значению 99,7%, которое можно было бы ожидать для нормального распределения.

е)



Гистограмма показывает, что значение 41 является возможным выбросом. Наличие такого значения смещает среднее в область высоких значений и увеличивает стандартное отклонение. Это может быть причиной того, что на расстоянии одного стандартного отклонения от среднего находится больше фирм, чем можно было бы ожидать, а именно 84,4%.

Упражнения с использованием базы данных

1. а) Размах равен $38 555.

б) Стандартное отклонение равно $10 806.

в) Коэффициент вариации равен 0,239, или 23,9%.

г) Разрыв между самым высоко- и самым низкооплачиваемым служащими составляет $38 555 (размах). Типичная разница между размером годовой заработной платы служащих и средним приблизительно равна $10 806 (стандартное отклонение), что составляет 23,9% от среднего. Размах больше, чем стандартное отклонение, поскольку размах измеряет наибольшую возможную разность между двумя значениями данных, а не типическое отличие от среднего.

Глава 6

Задачи

1. а) Случайный эксперимент заключается в следующем: вы ждете, пока будет объявлен размер доходов, и затем регистрируете это значение.

б) Выборка состоит из всего списка долларовых сумм, включая положительные, отрицательные и нулевые значения.

в) Результатом будут доходы компании Ford за прошлый квартал.

г) Список состоит из всех долларовых сумм, которые превышают вычисленное значение:

вычисленное значение + 0,01; вычисленное значение + 0,02; ...

д) Субъективная вероятность, поскольку она основана на мнении.

5. а) Вероятность равна 35/118 = 0,297.

б) Вероятность равна (1 –0,297) = 0,703.

6. а) Вероятность равна 0,22. Событие "большие затруднения" является дополнением события "А и В". Используя соотношение между *и* и *или*, находим:

вероятность (А и В) = 0,83 + 0,91 –0,96 = 0,78.

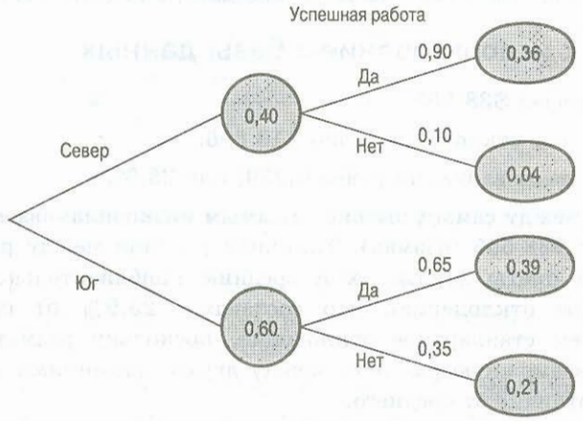
Используя правило дополнения, получаем:

вероятность "большие затруднения" = 1 –0,78 = 0,22.

б) Эти события не являются взаимно исключающими, поскольку вероятность события "А и В" равна 0,78 (п. "а").

в) Нет. Это не независимые события. Вероятность наступления обоих этих крайних сроков равна 0,78 (п. "а") и не равна произведению двух соответствующих вероятностей: 0,83 × 0,91 = 0,755. Такой же результат можно получить, используя условные вероятности.

10. а)



б) Вероятность успешной работы в первый год равна: 0,36 + 0,39 = 0,75.

в) Вероятность постройки в южной части и успешной работы равна 0,39.

г) Вероятность равна: 0,39/0,75 = 0,52. Это условная вероятность постройки на юге при условии успешной работы, которая равна: (вероятность "постройка на юге и успешная работа")/(вероятность "успешная работа").

д) Вероятность неуспеха работы при условии постройки в северной части равна: 0,04/0,4 = 0,10. Эта вероятность равна: (вероятность "неуспешная работа и постройка в северной части")/(вероятность "постройка в северной части").

Упражнения с использованием базы данных

1. а) Вероятность того, что будет выбрана женщина, равна: 28/71 = 0,394.

б) Вероятность того, что зарплата превышает $35 000, равна: 58/71 = 0,817.

в) Вероятность того, что зарплата превышает $35 000 при условии, что служащий имеет квалификацию В, равна: 0,310/0,338 = 0,917.

Глава 7

Задачи

1. а) Среднее значение платежа равно $8,50.

б) Ожидаемое значение платежа по опциону $8,50 показывает типическое или среднее значение случайного платежа.

в) Стандартное отклонение равно $10,14.

г) Стандартное отклонение, равное $10,14, характеризует степень риска для данного вложения капитала. Это обобщенная характеристика разности между действительным (случайным) размером выплаты и ожидаемым размером $8,50.

д) Вероятность равна: 0,15 + 0,10 = 0,25.

14. а) Мы допускаем, что все $n = 15$ ценных бумаг не зависят друг от друга и имеют одинаковые вероятности потерять стоимость.

б) Мы ожидаем, что 12 = (0,8) (15) ценных бумаг потеряют свою стоимость.

в) Стандартное отклонение равно: $\sigma_x = 1,55$.

г) Вероятность равна:

$$0,035 = \frac{15!}{15!\times 0!}0,8^{15}(1-0,8)^{0}$$ (или используйте таблицу).

д) Вероятность равна:

$$0,103 = \frac{15!}{10!\times 5!}0,8^{10}(1-0,8)^{5} = 3003\times 0,107374\times 0,00032$$

(или используйте таблицу).

24. а) Вероятность равна 0,75. Исходя из стандартизованного значения $z = (0,10-0,12)/0,03 = -0,67$ по стандартной таблице нормального распределения вероятностей для события "менее 10% рынка" находим значение 0,2514. Затем, используя правило дополнения, получаем ответ: 1 –0,2514 = 0,75.

Упражнения с использованием базы данных

1. б) $X = 52$; $p = 52/71 = 0,732$. Таким образом, 73,2% служащих имеют заработную плату свыше $40 000.

Глава 8

Задачи

1. а) Неприемлемо. Это нерепрезентативная выборка. Первым выпущенным за день трансмиссиям могли уделить больше внимания.

5. а) Статистика. Это среднее для рассматриваемой вами выборки.

   б) Параметр. Это среднее для всей генеральной совокупности.

8. Выборка состоит из счетов с номерами 43, 427 и 336. Последовательно берем по три случайные цифры и формируем числа 690, 043, 427, 336, 062,... . Первое число слишком большое (690 > 681, размер выборки), но следующие три можно использовать, и среди них нет повторов.

18. Вероятность равна 0,14. Стандартное отклонение среднего равно: $30/\sqrt{35} = 5{,}070926$, стандартизованные значения соответственно равны: $z_1 = (55-65)/5{,}070926 = -1{,}97$ и $z_2 = (60-65)/5{,}070926 = -0{,}99$. Для этих стандартизованных значений по таблице нормального распределения вероятностей находим соответствующие им значения и вычисляем разность: 0,1611 −0,0244 = 0,14.

22. а) Среднее равно: \$2601 × 45 = \$117045.

    б) Стандартное отклонение равно: $\$1275\sqrt{45} = \$8552{,}96$.

    в) В соответствии с центральной предельной теоремой.

    г) Вероятность равна 0,92. Вычисляем стандартизованное значение: $z = (105000-117045)/8552{,}96 = -1{,}41$. Для стандартизованного значения по стандартной таблице нормального распределения вероятностей находим число 0,0793, которое представляет вероятность того, что сумма заказа будет меньше \$105 000. Тогда вероятность того, что сумма заказа будет не менее \$105 000, равна: 1 −0,0793 = 0,92.

30. а) Стандартная ошибка среднего равна: $16{,}48/\sqrt{50} = \$2{,}33$. Это значение приблизительно показывает, насколько неизвестное значение среднего генеральной совокупности отличается от выборочного среднего \$53,01.

Упражнения с использованием базы данных

2. Объединив случайные цифры в группы по две, получим:

   14 53 62 38 70 78 40 24 17 59 26

   23 27 74 22 76 28 95 75

   Исключив числа больше 71 и меньше 1, получим:

   14 53 62 38 70 40 24 17 59 26

   23 27 22 28

   Первые 10 чисел не повторяются и дают такую выборку:

   14 53 62 38 70 40 24 17 59 26

   При желании их можно расположить в порядке возрастания номеров служащих:

   14 17 24 26 38 40 53 59 62 70

   а) Номера служащих: 14 53 62 38 70 40 24 17 59 26.

8. а) Значение биномиального X равно 5 женщинам.

   б) Стандартная ошибка составляет: $\sqrt{10\times0{,}5\times0{,}5} = 1{,}58$ и показывает, что наблюдаемое биномиальное распределенное X приблизительно на 1,58 больше или меньше среднего, которое вы могли бы ожидать в случайной выборке из 10 человек, взятых из той же генеральной совокупности.

Глава 9

Задачи

1. 95% доверительный интервал простирается от 101,26 до 105,54 бушеля на один акр (при t-значении 1,960). Доверительный интервал вычисляют по формуле:

   $103{,}6 - (1{,}960)\left(9{,}4/\sqrt{62}\right) = 103{,}6 - 2{,}34 = 101{,}26$ и
   $103{,}6 + (1{,}960)\left(9{,}4/\sqrt{62}\right) = 103{,}6 - 2{,}34 = 105{,}94$

   (Более точное вычисление на компьютере даст доверительный интервал от 101,21 до 105,99.)

5. а) 2,365; б) 3,499; в) 5,408; г) 1,895.

25. а) Среднее 2,34% характеризует доходность акций.

    б) Стандартное отклонение, равное 5,98%, показывает отклонение от среднего значения. Разность доходности типической акции из этого списка и среднего значения составляет около 5,98%.

    в) Стандартная ошибка, $5{,}979817/\sqrt{10} = 1{,}89\%$, показывает приблизительную разность (в процентных единицах) между средним выборки (2,34%) и неизвестным средним (идеализированной) генеральной совокупности аналогичных брокерских фирм.

    г) 95% доверительный интервал простирается от −1,94% до 6,62% (при t-значении 2,262).

    д) 90% доверительный интервал простирается от −1,13% до 5,81% (при t-значении 1,833). 90% двухсторонний доверительный интервал меньше, чем 95% двухсторонний доверительный интервал.

    е) Мы на 99% уверены, что средняя доходность акций из генеральной совокупности не меньше −2,99% (при t-значении 2,821).

    ж) Нет, вы должны использовать или тот же односторонний интервал, независимо от данных, или двухсторонний интервал. В противном случае вы можете не получить заявленный вами 99% доверительный интервал.

26. а) 95% доверительный интервал простирается от 49,2% до 55,6% (при t-значении 1,960). Доверительный интервал вычисляют по формуле:

$0,524 - 1,960\sqrt{0,524(1-0,524)/921} = 0,524 - 0,032 = 0,492$ и
$0,524 + 1,960\sqrt{0,524(1-0,524)/921} = 0,524 + 0,032 = 0,556$.

Упражнения с использованием базы данных

1. а) Среднее равно \$34 031,80. Стандартное отклонение равно \$10 472,93. Стандартная ошибка равна \$4683,64.

   б) 95% доверительный интервал простирается от \$21 030 до \$47 034 (при t-значении 2,776).

   в)



Глава 10

Задачи

1. а) Нулевая гипотеза $H_0 : \mu = 43,1$ утверждает, что среднее значение возраста в генеральной совокупности ваших клиентов равно среднему значению возраста жителей всего города. Исследуемая (альтернативная) гипотеза $H_1 : \mu \neq 43,1$ утверждает, что эти два значения различаются.

   б) Отклонить H_0 и принять H_1. Средний возраст клиентов значимо отличается от среднего возраста всего населения. 95% доверительный интервал простирается от 29,11 до 38,09 (при t-значении 1,960) и не включает заданное опорное значение (43,1). Значение t-статистики равно −4,15, что по абсолютной величине больше табличного значения, равного 2,576. (Более точные компьютерные вычисления дают следующие границы доверительного интервала: от 29,00 до 38,20.)

2. а) Отклонить H_0 и принять H_1. Средний возраст клиентов высоко значимо отличается от среднего возраста всего населения. 99% доверительный интервал простирается от 27,7 до 39,50 (при t-значении 1,960) и не включает заданное опорное значение (43,1). Значение t-статистики равно −4,15, что по абсолютной величине больше табличного значения 2,576. (Более точные компьютерные вычисления дают следующие границы доверительного интервала: от 27,46 до 39,74.)

   б) $p < 0,001$. t-значение для проверки гипотезы на уровне 0,001 равно 3,921.

36. а) Нет. Испытуемый под номером 4 имеет более высокое значение уровня стресса при правдивом ответе, чем при лживом.

   б) Средние значения уровней стресса равны: при правдивом ответе — 8,5, при лживом — 9,2. Разность средних значений равна 0,7.

в) Стандартная ошибка разности равна: $0,648074/\sqrt{6} = 0,264575$. Это зависимые выборки. Между двумя наборами данных существует естественная связь, поскольку в каждом наблюдении оба измерения проводились для одного и того же человека.

г) 95% двухсторонний доверительный интервал простирается от 0,02 до 1,38 (при t-значении 2,571). Мы на 95% уверены, что среднее значение разности уровней стресса (уровень стресса при лживом ответе минус уровень стресса при правдивом ответе) находится в пределах от 0,02 до 1,38.

д) Средние значения уровней стресса значимо отличаются ($p < 0,05$) между собой, поскольку опорное значение (0, свидетельствующее об отсутствии отличия) не находится в пределах доверительного интервала. Значение t-статистики равно: $0,7/0,264575 = 2,65$. Односторонний вывод для этого двухстороннего теста свидетельствует о том, что уровень стресса значимо выше при лживом ответе, чем при правдивом.

е) Уровень стресса значимо выше при лживом ответе, чем при правдивом. Это заключение об уровне стресса, сделанное на основании измерений уровня стресса у 6 человек — представителей генеральной совокупности, относится ко всей генеральной совокупности. Вывод сделан не для 6 человек, а для всей генеральной совокупности (реальной или идеальной), поскольку этих людей можно рассматривать как случайную выборку из этой генеральной совокупности. Хотя один человек представляет собой исключение (уровень его стресса при правдивом ответе выше, чем при лживом), заключение сделано относительно среднего значения разности уровней в генеральной совокупности. Однако это не означает, что данный вывод применим к каждому конкретному человеку.

38. а) Среднее время работы до поломки для вашего изделия равно 4,475 дня, для изделия вашего конкурента — 2,360 дня. Разность средних равна 2,115 дня.

   б) Стандартная ошибка равна 1,0066. Это две независимые выборки. Между сделанными измерениями нет естественной связи, поскольку они сделаны для различных объектов. Кроме того, эти выборки не связаны, поскольку два набора содержат разное количество измерений (разные размеры выборок).

   в) Двухсторонний 99% доверительный интервал простирается от −0,69 до 4,92 (при t-значении 2,787).

   г) Различие в надежности не значимо на уровне 1%. Опорное значение 0 находится внутри 99% доверительного интервала (п. в). Значение t-статистики равно 2,10.

   д) Различие в надежности значимо на уровне 5% ($p < 0,05$). Об этом свидетельствует 95% доверительный интервал (от 0,04 до 4,19), а также значение t-статистики (2,10), превышающее табличное t-значение (2,060) при проверке гипотезы на уровне значимости 5%. Согласно п. "г", результат проверки не значим на уровне 1%.

е) Проведенное исследование показало, что наши изделия значимо надежнее изделий наших конкурентов ($p < 0,05$, использовался t-тест для двух независимых выборок).

Упражнения с использованием базы данных

1. Да, среднее значение годовой заработной платы ($45 142) значимо отличается от $40 000, поскольку заданное опорное значение не находится в пределах 95% доверительного интервала (от $42 628 до $47 655 при t-значении 1,960). Кроме того, значение t-статистики равно:

   (45141,50 −40000)/1282,42 = 4,01.

   Отвергаем нулевую гипотезу и принимаем исследуемую (альтернативную) гипотезу о том, что среднее генеральной совокупности отличается от $40 000. (Более точный, вычисленный на компьютере, доверительный интервал составляет от $42 584 до $47 699.)

Глава 11

1. а)

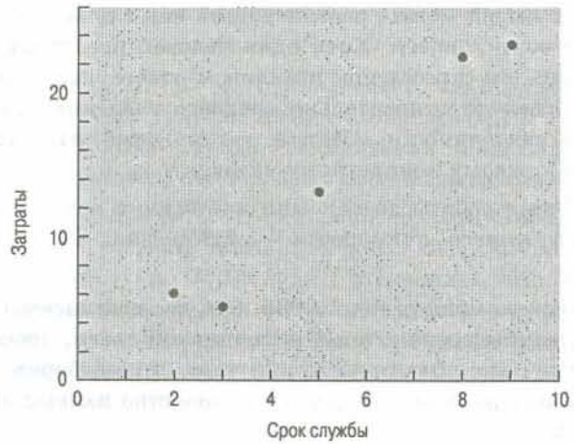
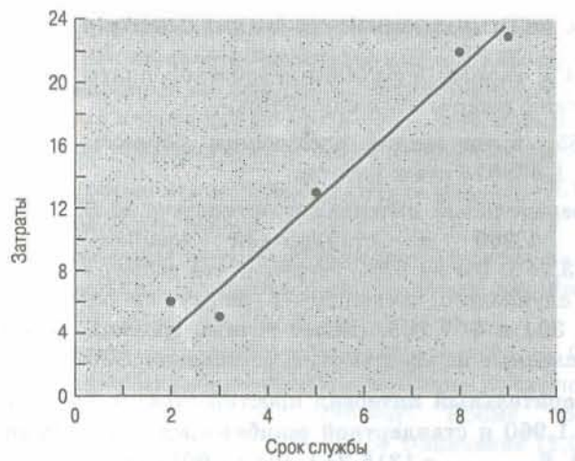


   Диаграмма рассеяния свидетельствует о наличии линейной структуры (положительная, возрастающая зависимость), значения данных распределены вокруг прямой линии, присутствует некоторая случайность.

   б) Коэффициент корреляции, r , между сроком службы пресса и стоимостью технического обслуживания равен 0,985. Значение коэффициента корреляции очень близок к 1, что указывает на сильную прямо пропорциональную (положительную) зависимость. Это согласуется с диаграммой рассеяния, на которой видно, что с увеличением срока службы пресса увеличивается (вдоль прямой линии) и стоимость его технического обслуживания.

   в) Прогнозируемая стоимость = −1,0645 + 2,7527 возраст.



г) Прогнозируемая стоимость = −1,06451 + (2,752688)(7) = 18,204 в тысячах долларов, т.е. $18 204.

д) $S_e = 1,7248$, в тысячах долларов, т.е. $1 725.

е) $R^2 = 96,9\%$ вариации стоимости обслуживания могут быть обусловлены тем, что одни прессы эксплуатируются дольше, чем другие.

ж) Да, срок эксплуатации объясняет значительную часть вариации стоимости ремонта. Это можно установить, проверив, действительно ли наклон прямой значимо отличен от нуля. Доверительный интервал для наклона (от 1,8527 до 3,6526) не включает 0, следовательно, наклон значимо отличается от нуля. Кроме того, заметим, что значение t-статистики, $t = b / S_b = 2,7527 / 0,2828 = 9,73$, превышает табличное t-значение (3,182) для 5 −2 = 3 степеней свободы, что также позволяет сделать вывод о значимости.

з) Дополнительная годовая стоимость обслуживания значимо отличается от $20 000. В соответствии с п. ж мы на 95% уверены в том, что долгосрочные издержки на ежегодное обслуживание одного пресса находятся где-то между $1853 и $3653 в год. Так как опорное значение $20 000 не находится в пределах доверительного интервала, то делаем вывод, что стоимость годового обслуживания одного печатного пресса значимо отличается от $20 000. Фактически эта сумма значимо меньше оценки $20 000, которую сделал ваш консервативно настроенный заместитель. Значение t-статистики равно: (2,752688 −20)/0,282790 = −61,0. Так как значение t-статистики больше критического значения (12,294 для 3 степеней свободы) на уровне значимости 0,001, то отвергаем нулевую гипотезу и принимаем исследуемую (альтернативную) гипотезу о том, что стоимость обслуживания печатного пресса для всей генеральной совокупности отличается от опорного значения, и утверждаем, что этот результат является очень высоко значимым ($p < 0,001$).

Упражнения с использованием базы данных

2. а) $R^2 = 30,4\%$ вариации размера заработной платы можно объяснить различием в стаже работы этих служащих.

   б) \$49 285, вычисленное следующим образом: прогноз зарплаты = 34575,94 + 1,838615 стаж работы.

   в) 95% доверительный интервал простирается от \$31 304 до \$67 265 (при t-значении 1,960 и стандартной ошибке нового наблюдения $S_{(Y|X_0)} = 9173,75$). Вы на 95% уверены, что размер годовой заработной платы нового служащего, имеющего 8 лет стажа работы, будет находиться между \$31 304 и \$67 265. (Более точные границы доверительного интервала, вычисленные на компьютере, составляют \$30 984 и \$67 586.)

   г) 95% доверительный интервал простирается от \$46 707 до \$51 863 (при t-значении 1,960 и стандартной ошибке среднего значения Y при данном X_0, равной $S_{(\text{прогноз. } Y|X_0)} = 1315,34$). Вы на 95% уверены, что средний размер годовой заработной платы для генеральной совокупности служащих, имеющих 8 лет стажа работы, находится между \$46 704 и \$51 863. (Более точные границы доверительного интервала, вычисленные на компьютере, равны \$46 661 и \$51 909.)

Глава 12

Задачи

1. Для определения прогнозируемого значения (Y = количество запросов в зависимости от X_1 = стоимость и X_2 = размер) используем множественную регрессию. Здесь лучше всего использовать F-тест, который проверяет значимость общего влияния всех переменных.

5. а) Цена = 8344,005 + 0,026260 (площадь) −4,26669 (год создания).

   б) Стоимость каждого дополнительного квадратного сантиметра площади картины равна \$26,26. При прочих равных условиях (например, для данного года создания) при увеличении площади на 1 квадратный сантиметр цена картины в среднем возрастает на (\$1000)(0,026260) = \$26,26.

   в) При постоянном значении площади картины коэффициент регрессии для года создания картины показывает, что с увеличением года создания картины на единицу ее стоимость уменьшается в среднем на \$4266,99. Картины, написанные раньше (более старые), имеют большую стоимость.

   г) Цена = 8344,005 + (0,026260) (4000) −(4,26669)(1954)=111,348 (тысяч) = \$111 348.

   д) Ошибка прогноза приблизительно равна \$153 111. Стандартная ошибка оценки, $S = 153,111$, показывает типичный размер ошибок прогноза для этого набора данных в тысячах долларов (поскольку значение Y выражено в тысячах).

   е) $R^2 = 28,2\%$ вариации цены картин Пикассо можно отнести на размер и год создания картины.

   ж) Да, регрессия является значимой. Результат F-теста, выполненного с использованием R^2, является значимым ($R^2 = 0,282$ для $n = 23$ наблюдений и $k = 2$ X-переменных превышает табличное значение 0,259). Это означает, что взятые вместе две переменные — площадь и год создания картины — объясняют значимую долю вариации цен на различные картины.

   Обычный F-тест также демонстрирует значимость, $F = 3,93203$, для 2 и 20 степеней свободы ($p = 0,036$).

   з) Да, площадь картины оказывает значимое влияние на ее цену, с учетом года создания картины. Большие полотна стоят значимо дороже созданных в этом же году меньших по размеру картин.

   Доверительный интервал простирается от 0,000896 до 0,051623. Доверительный интервал не включает опорное значение нуль, поэтому принимаем исследуемую (альтернативную) гипотезу. Кроме того, значение t-статистики (2,16) превышает критическое t-значение (2,086) для 20 степеней свободы ($p = 0,043$).

   и) Да, год создания картины оказывает значимое влияние на ее цену, с учетом размера полотна. Влияние года создания на цену заключается в уменьшении цены, т.е. более новые полотна стоят значимо меньше более старых картин того же размера. Доверительный интервал простирается от −8,0048 до −0,5291 и не включает опорное значение нуль. Кроме того, значение t-статистики −2,38 по абсолютной величине превышает критическое t-значение (2,086) ($p = 0,027$).

11. б) Прогнозируемая оплата = 583,3609 + 0,004369 (продажи) + 30,38801 (ROE) = 583,3609 + (0,004369)(77,721) + (30,38801)(15,0) = 1379 (тысяч) = \$1379000.

    Остаток = 1207 −1379 = −172 (тысячи) = −\$172000.

    Этому руководителю выплатили на \$172 000 меньше, чем можно было ожидать для фирмы с таким уровнем продаж и ROE.

Упражнения с использованием базы данных

1. а) Уравнение регрессии:

   зарплата = 22380,64 + 300,5515 (возраст) + 1579,259 (стаж работы).

   Это уравнение прогноза дает ожидаемый (в среднем) размер заработной платы типичного служащего данного возраста и с данным стажем работы. Каждый дополнительный год возраста добавляет к годовому размеру заработной платы в среднем \$300, а каждый дополнительный год трудового стажа оценивается в \$1579.

   б) $S_e = 8910,19$. Стандартная ошибка оценки показывает, что прогнозируемые размеры заработной платы отличаются от фактических приблизительно на \$8910.

   в) $R^2 = 0,3395$. Это означает, что 34% вариации размера заработной платы обусловлено возрастом служащих и трудовым стажем. Около 66% вариации обусловлено другими причинами.

г) Модель является значимой. Это свидетельствует о том, что возраст и стаж, взятые вместе, объясняют значимую долю вариации размера заработной платы. Для выполнения F-теста с помощью R^2 необходимо найти критическое значение по таблице. Поскольку $n = 71 > 50$, то для вычисления критического значения для уровня 5% используем множители (для $k = 2$ переменных):

Критическое значение $= 5,99/71 + (-0,27)/71^2 = 0,0843$.

Так как $R^2 = 0,3395$ превышает полученное критическое значение, то результат F-теста является значимым.

д) Возраст не оказывает значимого влияния на размер заработной платы, если не меняется стаж работы. Доверительный интервал (от −14 до 615) содержит опорное значение 0. Кроме того, значение t-статистики 1,91 меньше критического значения 1,960.

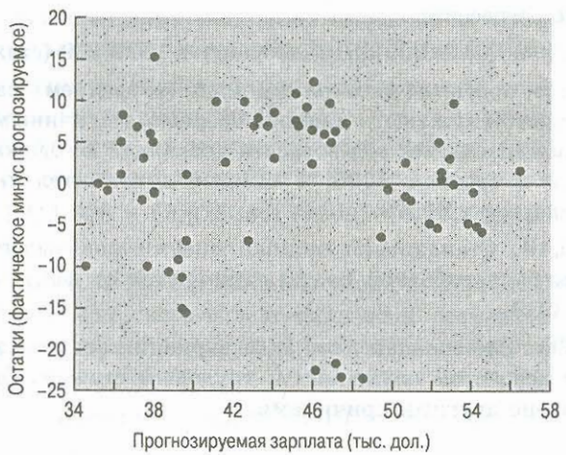
Стаж работы оказывает значительное влияние на размер зарплаты служащих, при условии, что возраст постоянен. Доверительный интервал для стажа работы (от 870 до 2289) не включает нуль. Значение t-статистики для стажа работы 4,44 превышает критическое значение 3,291 для уровня значимости 0,1%. Это свидетельствует о том, что стаж работы является высоко значимым в предсказании уровня заработной платы.

е) Стандартизованный коэффициент регрессии для возраста равен 0,203. Увеличение возраста на одно стандартное отклонение приводит к увеличению размера заработной платы на 20,35% стандартного отклонения.

Стандартизованный коэффициент регрессии для стажа работы равен 0,474. Увеличение стажа работы на одно стандартное отклонение приводит к увеличению размера заработной платы на 47,4% стандартного отклонения.

Это означает, что с точки зрения влияния на размер заработной платы стаж работы важнее, чем возраст, поскольку стандартизованный коэффициент регрессии для стажа работы выше. (Стандартные отклонения равны: 7,315 — для возраста, 3,241 — для стажа работы и 10,806 — для размера заработной платы.)

ж)



Ситуация выглядит как случайная: на диаграмме видна очень слабая (если вообще видна) структура. Однако есть группа из четырех значений в диапазоне от $46 000 до $49 000 (прогнозируемый размер зарплаты), которые имеют очень малые значения остатков в пределах от −20 до −25. Возможно, эту группу стоит изучать далее. В этой группе служащих, которым, вероятно, недоплачивают (поскольку остаток отрицательный) в соответствии с их возрастом и опытом работы, находятся только4 женщины с уровнем квалификации "А".

2. а) Прогноз зарплаты = 22 380,64 + (300,5515)(39) + (1579,259)(1) = $35 681.

   Ошибка прогноза = факт −прогноз = 35018 −35681 = −663.

   Прогнозируемый размер заработной платы ($35 681) приблизительно равен фактическому ($35 018). Ошибка прогноза (−663) предполагает, что заработная плата этих служащих на $663 меньше, чем можно было ожидать, исходя из их возраста и опыта работы.

Глава 13

Задачи

1. а) Цель: обеспечить предварительную информацию относительно возможностей других фирм для выбора стратегии экспансии. Аудитория: руководители, которым поручено планировать и принимать стратегические решения относительно этой экспансии.

2. а) Влияние. Как правило, в статистике используют существительное *эффект* ("этот (параметр) имеет эффект...") и глагол *влиять* ("этот параметр влияет...").

   б) Влияет.

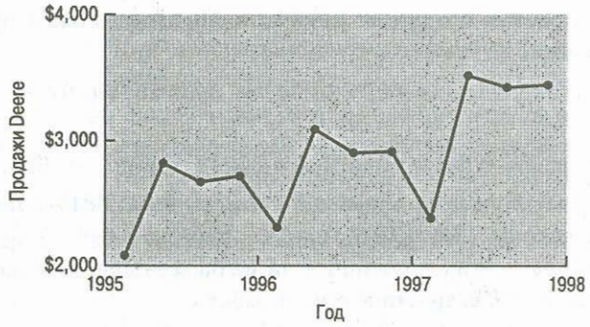
4. а) Анализ и методы, потому что это предложение содержит технические подробности и их интерпретацию.

5. а) White J. A. "How a Money Manager Can Pull a Rabbit Out of a Hat", *The Wall Street Journal*, 1989, March 16, p. C1.

6. а) Фамилия и должность человека, а также место, месяц и день.

Задачи

5. а)



Это в целом возрастающий с течением времени тренд, содержащий сезонные колебания.

б) Для первых двух кварталов 1995 г. скользящего среднего нет. Первое доступное значение для третьего квартала 1995 г.: (2088/2 +2673+ 2718 + 2318/2)/4 = $2602. Это и другие значения скользящего среднего приведены ниже.

Год	Продажи, млн дол.	Скользящее среднее
1995	2088	Отсутствует
1995	2812	Отсутствует
1995	2673	2602
1995	2718	2655
1996	2318	2729
1996	3089	2782
1996	2905	2817
1996	2917	2881
1997	2396	3000
1997	3521	3132
1997	3430	Отсутствует
1997	3444	Отсутствует



в) Сезонные индексы для четырех кварталов соответственно равны: 0,824; 1,117; 1,029 и 1,016. Да, эти значения выглядят обоснованными: индекс первого квартала всегда ниже соседних индексов.

г) Первый квартал наихудший. Объем продаж на 1 −0,824 = 17,6% ниже по сравнению с типичным кварталом.

д) Делим каждый из объемов продаж на соответствующий сезонный индекс.

Год	Продажи, млн дол.	С учетом сезонных колебаний
1995	2088	2534
1995	2812	2517
1995	2673	2597
1995	2718	2675
1996	2318	2813
1996	3089	2765
1996	2905	2822
1996	2917	2870
1997	2396	3908
1997	3521	3152
1997	3430	3332
1997	3444	3389

е) С учетом сезонных колебаний продажи также растут при переходе от третьего к четвертому кварталу 1995 г. (от 2597 до 2675).

ж) С учетом сезонных колебаний наблюдается рост продаж при переходе от второго к третьему кварталу 1997 г. (от 3152 до 3332).

з) Уравнение регрессии с использованием периода времени 1, 2, 3,... (переменная X) и ряда с поправкой на сезонность (переменная Y) имеет следующий вид:

прогноз продаж с поправкой = 2363 + 76,998 (период времени).

и) Прогноз продаж равен 4058 и вычислен путем подстановки 22 в качестве периода времени.

к) Прогноз равен 3534. Прогноз с поправкой на сезонность вычисляют на основе уравнения регрессии, используя период времени, равный 25, что позволяет определить прогноз продаж с поправкой как 4289. Внося сезонную поправку (путем умножения на сезонный индекс первого квартала), получаем прогноз: (4289)(0,824) = 3534.

8. б) 285 167/1,08 = 264 044.

   в) (264 043,5)(1,38) = 364 380.

9. а) 5423 + (17)(408) = 12 359.

   г) (12 359)(1,45) = 17 921.

Глава 15

Задачи

1. а) Наибольшую эффективность (68,1) имеет рекламный ролик №2, а наименьшую (53,5) — рекламный ролик №3.

   б) Общий объем выборки равен $n = 303$. Общее среднее равно $\bar{X} = 61,4073$. Количество выборок $k = 3$.

   в) Межгрупповая вариация равна 5617,30 с $k - 1 = 2$ степенями свободы.

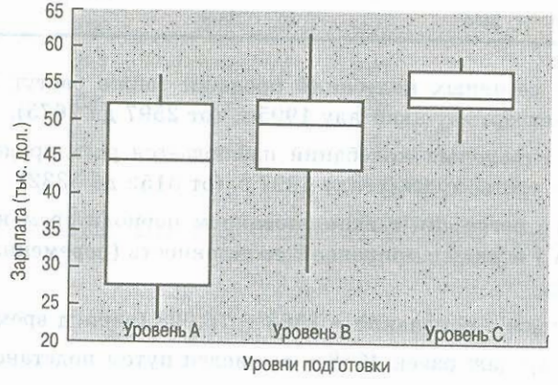
   г) Внутригрупповая вариация равна 91,006 с $n - k = 300$ степенями свободы.

4. б) Стандартная ошибка для разности средних равна 1,356192.

19. а) Да. Разность в производительности между сотрудничающей и конкурирующей группами является статистически значимой. Производительность в сотрудничающей группе значимо выше, чем в конкурирующей. Эта разница значима, поскольку p-значение 0,049682 (для теста "Соревнование — сотрудничество (А)" в таблице) меньше 0,05, что свидетельствует о значимости на уровне 5%.

Упражнения с использованием базы данных

1. а)



Видно, что, как правило, более высокую зарплату имеют служащие с более высоким уровнем подготовки. Самые высокие зарплаты в группе С, в которую входят служащие, прошедшие больше одного курса подготовки, а самые низкие зарплаты имеют служащие из группы А, в которую входят сотрудники, не прошедшие ни одного курса подготовки. В целом оклады выше в группах В и С, т.е. у служащих, прошедших курсы обучения.

Группы характеризуются различным уровнем вариации: наибольший — в группе А и наименьший — в группе С.

б) Средние равны: в группе А — $41 010,87; в группе В — $48 387,17; в группе С — $53 926,89. Среднее значение оклада увеличивается при увеличении уровня подготовки. Это похоже на то, что наблюдается для значений медианы на блочных диаграммах.

в) Межгрупповая вариация равна 797 916 214 с $k - 1 = 2$ степенями свободы.

Внутригрупповая вариация 96 732 651 с $n - k = 68$ степенями свободы.

Глава 16

Задачи

2. г) На основании результатов непараметрического теста отвергаем нулевую гипотезу. Медиана прибыли фирм, занимающихся строительными материалами, значимо отличается от убытков в размере 5 процентных единиц. Ниже приведены все шаги теста.

   1. Размер модифицированной выборки равен 14. Ни одно из значений данных не равно опорному значению −5.
   2. По таблице определяем, что результат критерия знаков является значимым на уровне 5%, если количество таких порядковых значений, которые меньше опорного, меньше 3 или больше 11.
   3. Подсчет показывает, что в наборе данных два значения лежат ниже опорного значения −5. Это фирмы Manville (значение −59) и National Gypsum (значение −7).
   4. Полученное количество лежит вне границ (для уровне значимости 5%), найденных по таблице для критерия знаков. Поэтому можно утверждать, что медиана прибыли фирм, занимающихся строительными материалами, значимо отличается от убытков в размере 5 процентных единиц.

6. б) Размер модифицированной выборки, количество тех, у кого состояние изменилось, составляет 8 + 2 = 10.

   в) Принимаем нулевую гипотезу; разница не является значимой. Ниже приведены подробности проверки.

   Размер модифицированной выборки равен 10.

   1. Из таблицы определяем, что результат критерия знаков является значимым на уровне 5%, если количество таких порядковых значений, которые меньше опорного, меньше 2 или больше 8.
   2. Подсчет показывает, что есть 2 значения, которые меньше опорного, и 8 значений, больших опорного.

3. Поскольку полученное количество находится внутри границ, найденных по таблице для критерия знаков, то различие не является статистически значимым.

10. б) Принимаем исследуемую (альтернативную) гипотезу, поскольку значение тестовой статистики 4,53 превышает 1,960. Различие в заработной плате между мужчинами и женщинами значимо при использовании непараметрического теста для двух независимых выборок. Ниже приведены детали. Сначала представлена таблица, содержащая все значения с их рангами (ранги усреднены, где это необходимо).

Ранг	Зарплата, дол.	Пол
1	20700	Ж
2	20900	Ж
3	21100	Ж
4	21900	Ж
5	22800	Ж
6	23000	Ж
7	23100	Ж
8	24700	Ж
9	25000	Ж
10	25500	М
11	25800	Ж
12	26100	М
13,5	26200	Ж
13,5	26200	М
15	26900	Ж
16	27100	Ж
17	27300	М
18	28100	Ж
19	29100	М
20	29700	Ж
21	30300	М
22	30700	М
23	32100	М
24	32800	М
25	33300	М
26,5	34000	М
26,5	34000	М
28	34100	М
30	35700	М
30	35700	М

Ранг	Зарплата, дол.	Пол
30	35700	М
32	35800	М
33	36900	М
34	37400	М
35	38100	М
36	38600	М
37	38700	М

Разделив данные на две группы, получим.

Ранг	Зарплата, дол.	Пол
1	20700	Ж
2	20900	Ж
3	21100	Ж
4	21900	Ж
5	22800	Ж
6	23000	Ж
7	23100	Ж
8	24700	Ж
9	25000	Ж
11	25800	Ж
13,5	26200	Ж
15	26900	Ж
16	27100	Ж
18	28100	Ж
20	29700	Ж

Ранг	Зарплата, дол.	Пол
10	25500	М
12	26100	М
13,5	26200	М
17	27300	М
19	29100	М
21	30300	М
22	30700	М
23	32100	М
24	32800	М
25	33300	М

Окончание таблицы

Ранг	Зарплата, дол.	Пол
26,5	34000	М
26,5	34000	М
28	34100	М
30	35700	М
30	35700	М
30	35700	М
32	35800	М
33	36900	М
34	37400	М
35	38100	М
36	38600	М
37	38700	М

Средний ранг для служащих-женщин равен 9,23333, средний ранг для служащих-мужчин — 25,65909, разность средних рангов — 16,42576, стандартная ошибка — 3,624481, значение тестовой статистики — 4,53.

Глава 17

Задачи

2. в) Вы ожидаете 247,63, или 248 человек. Это так, поскольку 46,2% из 536 человек предпочтут малолитражный автомобиль: $0{,}462 \times 536 = 247{,}63$.

   г) Для вычисления ожидаемого количества людей умножаем опорное значение доли в совокупности на размер выборки. Для семейного седана опорное значение доли равно 0,258, а размер выборки у нас равен 536. Отсюда ожидаемое количество людей равно: $0{,}258 \times 536 = 138{,}288$.

Тип машины	Процент предпочтений в прошлом году	Ожидаемое количество людей
Семейный седан	25,8	138,288
Малолитражный автомобиль	46,2	247,632
Спортивный автомобиль	8,1	43,416
Автомобиль-фургон	12,4	66,464
Пикап	7,5	40,200
Итого	100	536

   д) Значение "хи-квадрат" статистики равно 29,49. Ниже приведены наблюдаемые и ожидаемые частоты.

   "Хи-квадрат" $= (187 - 138{,}288)^2/138{,}288 + (206 - 247{,}632)^2/247{,}632 +$
   $+ (29 - 43{,}416)^2/43{,}416 + (72 - 66{,}464)^2/66{,}464 + (42 - 40{,}200)^2/40{,}200 =$
   $= 17{,}159 + 6{,}999 + 4{,}787 + 0{,}461 + 0{,}081 = 29{,}49.$

Тип	Частота, наблюдаемая на прошлой неделе	Ожидаемая частота
Семейный седан	187	138,288
Малолитражный автомобиль	206	247,632
Спортивный автомобиль	29	43,416
Автомобиль-фургон	72	66,464
Пикап	42	40,200
Итого	536	536

6. г) "Хи-квадрат" равен 5,224.

Глава 18

Задачи

1. а) Диаграмма Парето. Диаграмма Парето показывает проблемы, упорядоченные от наиболее часто встречающихся к наименее часто встречающимся, что позволяет сфокусировать внимание на наиболее важных проблемах.

   б) R-карта. R-карта дает возможность контролировать вариабельность процесса, чтобы при необходимости корректировать его. В данном случае мы имеем проблему, поскольку изготавливаемые двигатели отличаются друг от друга.

6. а) Центральная линия: $\overline{\overline{X}} = 56{,}31$. Контрольные границы вычисляются по формулам $\overline{\overline{X}} - A_2\overline{R}$ и $\overline{\overline{X}} + A_2\overline{R}$, где $A_2 = 0{,}483$, и простираются от 54,30 до 58,32.

7. б) $\overline{\overline{X}} = 12{,}8423$, $\overline{R} = 0{,}208$.

   в) Центральная линия равна 12,8423.

   г) Контрольные границы вычисляются по формулам $\overline{\overline{X}} - A_2\overline{R}$ и $\overline{\overline{X}} + A_2\overline{R}$, где $A_2 = 1{,}023$, и простираются от 12,630 до 13,055.

8. а) Центральная линия равна: $\overline{R} = 0{,}208$

   б) Контрольные границы вычисляются по формулам $D_3\overline{R}$ и $D_4\overline{R}$, где $D_3 = 0$ и $D_4 = 2{,}574$, и простираются от 0 до 0,535.

11. а) Центральная линия равна: $\overline{p} = 0{,}0731$. Контрольные границы простираются от 2,80% до 11,82%, что вычисляется по формулам

$$\overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \text{ и } \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}.$$

   в) Центральная линия равна: $\pi_0 = 0{,}0350$. Контрольные границы простираются от 1,55% до 5,45%, что вычисляется по формулам

$$\pi_0 - 3\sqrt{\frac{\pi_0(1-\pi_0)}{n}} \text{ и } \pi_0 + 3\sqrt{\frac{\pi_0(1-\pi_0)}{n}}.$$

Статистические таблицы

Таблица

В.1. Стандартная таблица нормального распределения вероятностей

В.2. Таблица случайных чисел

В.3. Таблица биномиального распределения вероятностей.

В.4. t-таблица.

В.5. R^2-таблица: критические значения для 5% уровня значимости (значимый результат).

В.6. R^2-таблица: критические значения для 1% уровня значимости (высоко значимый результат).

В.7. R^2-таблица: критические значения для 0,1% уровня значимости (очень высоко значимый результат).

В.8. R^2-таблица: критические значения для 10% уровня значимости.

В.9. F-таблица: критические значения для 5% уровня значимости (значимый результат).

В.10. F-таблица: критические значения для 1% уровня значимости (высоко значимый результат).

В.11. F-таблица: критические значения для 0,1% уровня значимости (очень высоко значимый результат).

В.12. F-таблица: критические значения для 10% уровня значимости

В.13. Ранги для критерия знаков.

В.14. Критические значения для тестов хи-квадрат.

В.15. Множители для построения $\bar{X}$-и R-карт.

Таблица В.1. Стандартная таблица нормального распределения вероятностей (см. рис. 7.3.5)

z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность
−2,00	0,0228	−1,00	0,1587	0,00	0,5000	0,00	0,5000	1,00	0,8413	2,00	0,9772
−2,01	0,0222	−1,01	0,1562	−0,01	0,4960	0,01	0,5040	1,01	0,8438	2,01	0,9778
−2,02	0,0217	−1,02	0,1539	−0,02	0,4920	0,02	0,5080	1,02	0,8461	2,02	0,9783
−2,03	0,012	−1,03	0,1515	−0,03	0,4880	0,03	0,5120	1,03	0,8485	2,03	0,9788
−2,04	0,0207	−1,04	0,1492	−0,04	0,4840	0,04	0,5160	1,04	0,8508	2,04	0,9793
−2,05	0,0202	−1,05	0,1469	−0,05	0,4801	0,05	0,5199	1,05	0,8531	2,05	0,9798
−2,06	0,0197	−1,06	0,1446	−0,06	0,4761	0,06	0,5239	1,06	0,8554	2.06	0,9803
−2,07	0,0192	−1,07	0,1423	−0,07	0,4721	0,07	0,5279	1.07	0,8577	2,07	0,9808
−2,08	0,0188	−1,08	0,1401	−0,08	0,4681	0,08	0,5319	1,08	0,8599	2,08	0,9812
−2,09	0,0183	−1,09	0,1379	−0,09	0,4641	0,09	0,5359	1,09	0,8621	2,09	0,9817
−2,10	0,0179	−1,10	0,1357	−0,10	0,4602	0,10	0,5398	1,10	0,8643	2,10	0,9821
−2,11	0,0174	−1,11	0,1335	−0,11	0,4562	0,11	0,5438	1,11	0,8665	2,11	0,9826
−2,12	0,0170	−1,12	0,1314	−0,12	0,4522	0,12	0,5478	1,12	0,8686	2,12	0,9830
−2,13	0,0166	−1,13	0,1292	−0,13	0,4483	0,13	0,5517	1,13	0,8708	2,13	0,9834
−2,14	0,0162	−1,14	0,1271	−0,14	0,4443	0,14	0,5557	1,14	0,8729	2,14	0,9838
−2,15	0,0158	−1,15	0,1251	−0,15	0,4404	0,15	0,5596	1,15	0,8749	2,15	0,9842
−2,16	0,0154	−1,16	0,1230	−0,16	0,4364	0,16	0,5636	1,16	0,8770	2,16	0,9846
−2,17	0,0150	−1,17	0,1210	−0,17	0,4325	0,17	0,5675	1,17	0,8790	2,17	0,9850
−2,18	0,0146	−1,18	0,1190	−0,18	0,4286	0,18	0,5714	1,18	0,8810	2,18	0,9854
−2,19	0,0143	−1,19	0,1170	−0,19	0,4247	0,19	0,5753	1,19	0,8830	2,19	0,9857
−2,20	0,0139	−1,20	0,1151	−0,20	0,4207	0,20	0,5793	1,20	0,8849	2,20	0,9861
−2,21	0,0136	−1,21	0,1131	−0,21	0,4168	0,21	0,5832	1,21	0,8869	2,21	0,9864
−2,22	0,0132	−1,22	0,1112	−0,22	0,4129	0,22	0,5871	1,22	0,8888	2,22	0,9868
−2,23	0,0129	−1,23	0,1093	−0,23	0,4090	0,23	0,5910	1,23	0,8907	2,23	0,9871
−2,24	0,0125	−1,24	0,1075	−0,24	0,4052	0,24	0,5948	1,24	0,8925	2,24	0,9875
−2,25	0,0122	−1,25	0,1056	−0,25	0,4013	0,25	0,5987	1,25	0,8944	2,25	0,9878
−2,26	0,0119	−1,26	0,1038	−0,26	0,3974	0,26	0,6026	1,26	0,8962	2,26	0,9881
−2,27	0,0116	−1,27	0,1020	−0,27	0,3936	0,27	0,6064	1,27	0,8980	2,27	0,9884
−2,28	0,0113	−1,28	0,1003	−0,28	0,3897	0,28	0,6103	1,28	0,8997	2,28	0,9887
−2,29	0,0110	−1,29	0,0985	−0,29	0,3859	0,29	0,6141	1,29	0,9015	2,29	0,9890
−2,30	0,0107	−1,30	0,0968	−0,30	0,3821	0,30	0,6179	1,30	0,9032	2,30	0,9893
−2,31	0,0104	−1,31	0,0951	−0,31	0,3783	0,31	0,5217	1,31	0,9049	2,31	0,9896

z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность
–2,32	0,0102	–1,32	0,0934	–0,32	0,3745	0,32	0,6255	1,32	0,9066	2,32	0,9898
–2,33	0,0099	–1,33	0,0918	–0,33	0,3707	0,33	0,6293	1,33	0,9082	2,33	0,9901
–2,34	0,0096	–1,34	0,0901	–0,34	0,3669	0,34	0,6331	1,34	0,9099	2,34	0,9904
–2,35	0,0094	–1,35	0,0885	–0,35	0,3632	0,35	0,6368	1,35	0,9115	2,35	0,9906
–2,36	0,0091	–1,36	0,0869	–0,36	0,3594	0,36	0,6406	1,36	0,9131	2,36	0,9909
–2,37	0,0089	–1,37	0,853	–0,37	0,3557	0,37	0,6443	1,37	0,9147	2,37	0,9911
–2,38	0,0087	–1,38	0,838	–0,38	0,3520	0,38	0,6480	1,38	0,9162	2,38	0,9913
–2,39	0,0084	–1,39	0,823	–0,39	0,3483	0,39	0,6517	1,39	0,9177	2,39	0,9916
–2,40	0,0082	–1,40	0,808	–0,40	0,3446	0,40	0,6554	1,40	0,9192	2,40	0,9918
–2,41	0,0080	–1,41	0,0793	–0,41	0,3409	0,41	0,6591	1,41	0,9207	2,41	0,9920
–2,42	0,0078	–1,42	0,0778	–0,42	0,3372	0,42	0,6628	1,42	0,9222	2,42	0,9922
–2,43	0,0075	–1,43	0,0764	–0,43	0,3336	0,43	0,6664	1,43	0,9236	2,43	0,9925
–2,44	0,0073	–1,44	0,0749	–0,44	0,3300	0,44	0,6700	1,44	0,9251	2,44	0,9927
–2,45	0,0071	–1,45	0,0735	–0,45	0,3264	0,45	0,6736	1,45	0,9265	2,45	0,9929
–2,46	0,0069	–1,46	0,0721	–0,46	0,3228	0,46	0,6772	1,46	0,9279	2,46	0,9931
–2,47	0,0068	–1,47	0,0708	–0,47	0,3192	0,47	0,6808	1,47	0,9292	2,47	0,9932
–2,48	0,0066	–1,48	0,0694	–0,48	0,3156	0,48	0,6844	1,48	0,9306	2,48	0,9934
–2,49	0,0064	–1,49	0,0681	–0,49	0,3121	0,49	6879	1,49	0,9319	2,49	0,9936
–2,50	0,0062	–1,50	0,0668	–0,50	0,3085	0,50	0,6915	1,50	0,9332	2,50	0,9938
–2,51	0,0060	–1,51	0,0655	–0,51	0,3050	0,51	0,6950	1,51	0,9345	2,51	0,9940
–2,52	0,0059	–1,52	0,0643	–0,52	0,3015	0,52	0,6985	1,52	0,9357	2,52	0,9941
–2,53	0,0057	–1,53	0,0630	–0,53	0,2981	0,53	0,7019	1,53	0,9370	2,53	0,9943
–2,54	0,0055	–1,54	0,0618	–0,54	0,2946	0,54	0,7054	1,54	0,9382	2,54	0,9945
–2,55	0,0054	–1,55	0,0606	–0,55	0,2912	0,55	0,7088	1,55	0,9394	2,55	0,9946
–2,56	0,0053	–1,56	0,0594	–0,56	0,2877	0,56	0,7123	1,56	0,9406	2,56	0,9948
–2,57	0,0051	–1,57	0,0582	–0,57	0,2843	0,57	0,7157	1,57	0,9418	2,57	0,9949
–2,58	0,0049	–1,58	0,0571	–0,58	0,2810	0,58	0,7190	1,58	0,9429	2,58	0,9951
–2,59	0,0048	–1,59	0,0559	–0,59	0,2776	0,59	0,7224	1,59	0,9441	2,59	0,9952
–2,60	0,0047	–1,60	0,0548	–0,60	0,2743	0,60	0,7257	1,60	0,9452	2,60	0,9953
–2,61	0,0045	–1,61	0,0537	–0,61	0,2709	0,61	0,7291	1,61	0,9463	2,61	0,9955
–2,62	0,0044	–1,62	0,0526	–0,62	0,2676	0,62	0,7324	1,62	0,9474	2,62	0,9956
–2,63	0,0043	–1,63	0,0516	–0,63	0,2643	0,63	0,7357	1,63	0,9484	2,63	0,9957

z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность
–2,64	0,0041	–1,64	0,0505	–0,64	0,2611	0,64	0,7389	1,64	0,9495	2,64	0,9959
–2,65	0,0040	–1,65	0,0495	–0,65	0,2578	0,65	0,7422	1,65	0,9505	2,65	0,9960
–2,66	0,0039	–1,66	0,0485	–0,66	0,2546	0,66	0,7454	1,66	0,9515	2,66	0,9961
–2,67	0,0038	–1,67	0,0475	–0,67	0,2514	0,67	0,7486	1,67	0,9525	2,67	0,9962
–2,68	0,0037	–1,68	0,0465	–0,68	0,2483	0,68	0,7517	1,68	0,9535	2,68	0,9963
–2,69	0,0036	–1,69	0,0455	–0,69	0,2451	0,69	0,7549	1,69	0,9545	2,69	0,9964
–2,70	0,0035	–1,70	0,0446	–0,70	0,2420	0,70	0,7580	1,70	0,9554	2,70	0,9965
–2,71	0,0034	–1,71	0,0436	–0,71	0,2389	0,71	0,7611	1,71	0,9564	2,71	0,9966
–2,72	0,0033	–1,72	0,0427	–0,72	0,2358	0,72	0,7642	1,72	0,9573	2,72	0,9967
–2,73	0,0032	–1,73	0,0418	–0,73	0,2327	0,73	0,7673	1,73	0,9582	2,73	0,9968
–2,74	0,0031	–1,74	0,0409	–0,74	0,2296	0,74	0,7704	1,74	0,9591	2,74	0,9969
–2,75	0,0030	–1,75	0,0401	–0,75	0,2266	0,75	0,7734	1,75	0,9599	2,75	0,9970
–2,76	0,0029	–1,76	0,0392	–0,76	0,2236	0,76	0,7764	1,76	0,9608	2,76	0,9971
–2,77	0,0028	–1,77	0,0384	–0,77	0,2206	0,77	0,7794	1,77	0,9616	2,77	0,9972
–2,78	0,0027	–1,78	0,0375	–0,78	0,2177	0,78	0,7823	1,78	0,9625	2,78	0,9973
–2,79	0,0026	–1,79	0,0367	–0,78	0,2148	0,79	0,7852	1,79	0,9633	2,79	0,9974
–2,80	0,0026	–1,80	0,0359	–0,80	0,2129	0,80	0,7881	1,80	0,9641	2,80	0,9974
–2,81	0,0025	–1,81	0,0351	–0,81	0,2090	0,81	0,7910	1,81	0,9649	2,81	0,9975
–2,82	0,0024	–1,82	0,0344	–0,82	0,2061	0,82	0,7939	1,82	0,9656	2,82	0,9976
–2,83	0,0023	–1,83	0,0336	–0,83	0,2033	0,83	0,7967	1,83	0,9664	2,83	0,9977
–2,84	0,0023	–1,84	0,0329	–0,84	0,2005	0,84	0,7995	1,84	0,9671	2,84	0,9977
–2,85	0,0022	–1,85	0,0322	–0,85	0,1977	0,85	0,8023	1,85	0,9678	2,85	0,9978
–2,86	0,0021	–1,86	0,0314	–0,86	0,1949	0,86	0,8051	1,86	0,9686	2,86	0,9979
–2,87	0,0021	–1,87	0,0307	–0,87	0,1922	0,87	0,8078	1,87	0,9693	2,87	0,9979
–2,88	0,0020	–1,88	0,0301	–0,88	0,1894	0,88	0,8106	1,88	0,9699	2,88	0,9980
–2,89	0,0019	–1,89	0,0294	–0,89	0,1867	0,89	0,8133	1,89	0,9706	2,89	0,9981
–2,90	0,0019	–1,90	0,0287	–0,90	0,1841	0,90	0,8159	1,90	0,9713	2,90	0,9981
–2,91	0,0018	–1,91	0,0281	–0,91	0,1814	0,91	0,8186	1,91	0,9719	2,91	0,9982
–2,92	0,0018	–1,92	0,0274	–0,92	0,1788	0,92	0,8212	1,92	0,9736	2,92	0,9982
–2,93	0,0017	–1,93	0,0268	–0,93	0,1762	0,93	0,8238	1,93	0,9732	2,93	0,9983
–2,94	0,0016	–1,94	0,0262	–0,94	0,1736	0,94	0,8264	1,94	0,9738	2,94	0,9984
–2,95	0,0016	–1,95	0,0256	–0,95	0,1711	0,95	0,8289	1,95	0,9744	2,95	0,9984

z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность	z-значение	Вероятность
–2,96	0,0015	–1,96	0,0250	–0,96	0,1685	0,96	0,8315	1,96	0,9750	2,96	0,9985
–2,97	0,0015	–1,97	0,0244	–0,97	0,1660	0,97	0,8340	1,97	0,9756	2,97	0,9985
–2,98	0,0014	–1,98	0,0239	–0,98	0,1635	0,98	0,8365	1,98	0,9761	2,98	0,9986
–2,99	0,0014	–1,99	0,0233	–0,99	0,1611	0,99	0,8389	1,99	0,9767	2,99	0,9986
–3,00	0,0013	–2,00	0,0228	–1,00	0,1587	1,00	0,8413	2,00	0,9772	3,00	0,9987

Таблица В.2. Таблица случайных чисел

	1	2	3	4	5	6	7	8	9	10
1	51449	39284	85527	67168	91284	19954	91166	70918	85957	19492
2	16144	56830	67507	97275	25982	69294	32841	20861	83114	12531
3	48145	48280	99481	13050	81818	25282	66466	24461	97021	21072
4	83780	48351	85422	42978	26088	17869	94245	26622	48318	73850
5	95329	38482	93510	39170	63683	40587	80451	43058	81923	97072
6	11179	69004	34273	36062	26234	58601	47159	82248	95968	99722
7	94631	52413	31524	02316	27611	15888	13525	43809	40014	30667
8	64275	10294	35027	25604	65695	36014	17988	02734	31732	29911
9	72125	19232	10782	30615	42005	90419	32447	53688	36125	28456
10	16463	42028	27927	48403	88963	79615	41218	43290	53618	68082
11	10036	66273	69506	19610	01479	92338	55140	81097	73071	61544
12	85356	51400	88502	98267	73943	25828	38219	13268	09016	77465
13	84076	82087	55053	75370	71030	92275	55497	97123	40919	57479
14	76731	39755	78537	51937	11680	78820	50082	56068	36908	55399
15	19032	73472	79399	05549	14772	32746	38841	45524	13535	03113
16	72791	59040	61529	74437	74482	76619	05232	28616	98690	24011
17	11553	00135	28306	65571	34465	47423	39198	54456	95283	54637
18	71405	70352	46763	64002	62461	41982	15933	46942	36941	93412
19	17594	10116	55483	96219	85493	96955	89180	59690	82170	77643
20	09584	23476	09243	65568	89128	36747	63692	09986	47687	46448

	1	2	3	4	5	6	7	8	9	10
21	81667	62634	52794	01466	85938	14565	79993	44956	82254	65223
22	45849	01177	13773	43523	69825	03222	58458	77463	58521	07273
23	97252	92257	90419	01241	52516	66293	14536	23870	78402	41759
24	26232	77422	76289	57587	42831	87047	20092	92676	12017	43554
25	87779	33602	01931	66913	63008	03745	93939	07178	70003	18158
26	46120	62298	69126	07862	76731	58527	39342	42749	57050	91725
27	53292	55652	11834	47581	25682	64085	26587	92289	41853	38354
28	81606	56009	06021	98392	40450	87721	50917	16978	39472	23505
29	67819	47314	96988	89931	49395	37071	72658	53947	11996	64631
30	50458	20350	87362	83996	85422	58694	71813	97695	28804	58523
31	59772	27000	97805	25042	09916	77569	71347	62667	09330	02152
32	94752	91056	08939	93410	59204	04644	44336	55570	21106	76588
33	01885	82054	45944	55398	55487	56455	56940	68787	36591	29914
34	85190	91941	86714	76593	77199	39724	99548	13827	84961	76740
35	97747	67607	14549	08215	95408	46381	12449	03672	40325	77312
36	43318	84469	26047	86003	34786	38931	34846	28711	42833	93019
37	47874	71365	76603	57440	49514	17335	71969	58055	99136	73589
38	24259	48079	71198	95859	94212	55402	93392	31965	94622	11673
39	31947	64805	34133	03245	24546	48934	41730	47831	26531	02203
40	37911	93224	87153	54541	57529	38299	65659	00202	07054	40168
41	82714	15799	93126	74180	94171	97117	31431	00323	62793	11995
42	82927	37884	74411	45887	36713	52339	68421	35968	67714	05883
43	65934	21782	35804	36676	35404	69987	52268	19894	81977	87764
44	56953	04356	68903	21369	35901	86797	83901	68681	02397	55359
45	16278	17165	67843	49349	90163	97337	35003	34915	91485	33814
46	96339	95028	48468	12279	81039	56531	10759	19579	00015	22829
47	84110	49661	13988	75909	35580	18426	29038	79111	56049	96451
48	49017	60748	03412	09880	94091	90052	43596	21424	16584	67970
49	43560	05552	54344	69418	01327	07771	25364	77373	34841	75927
50	25206	15177	63049	12464	16149	18759	96184	15968	89446	07168

Таблица В.3. Таблица биномиального распределения вероятностей. Точное значение вероятностей находится в колонке под заголовком "Точное", кумулятивное значение — в колонке под заголовком "Сумма". Например, вероятность того, что биномиально распределенная случайная переменная при $\pi = 0{,}30$ и $n = 3$ точно равна $a = 2$, определяется из таблицы как 0,189. Вероятность того, что эта биномиально распределенная случайная переменная меньше или равна $a = 2$, определяется как 0,973 (сумма вероятностей для $a = 0$, 1, и 2)

		$\pi = 0{,}05$		$\pi = 0{,}10$		$\pi = 0{,}20$		$\pi = 0{,}30$		$\pi = 0{,}40$		$\pi = 0{,}50$		$\pi = 0{,}60$		$\pi = 0{,}70$		$\pi = 0{,}80$		$\pi = 0{,}90$		$\pi = 0{,}95$	
n	*a*	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
1	0	0,950	0,950	0,900	0,900	0,800	0,800	0,700	0,700	0,600	0,600	0,500	0,500	0,400	0,400	0,300	0,300	0,200	0,200	0,100	0,100	0,050	0,050
1	1	0,050	1,00	0,100	1,00	0,200	1,000	0,300	1,000	0,400	1,000	0,500	1,000	0,600	1,000	0,700	1,000	0,800	1,000	0,900	1,000	0,950	1,000
2	0	0,903	0,903	0,810	0,810	0,640	0,640	0,490	0,490	0,360	0,360	0,250	0,250	0,160	0,160	0,090	0,090	0,040	0,040	0,010	0,010	0,003	0,003
2	1	0,095	0,998	0,180	0,990	0,320	0,960	0,420	0,910	0,480	0,840	0,500	0,750	0,480	0,640	0,420	0,510	0,320	0,360	0,180	0,190	0,095	0,098
2	2	0,003	1,000	0,010	1,000	0,040	1,000	0,090	1,000	0,160	1,000	0,250	1,000	0,360	1,000	0,490	1,000	0,640	1,000	0,810	1,000	0,903	1,000
3	0	0,857	0,857	0,729	0,729	0,512	0,512	0,343	0,343	0,216	0,216	0,125	0,125	0,064	0,064	0,027	0,027	0,008	0,008	0,001	0,001	0,000	0,000
3	1	0,135	0,993	0,243	0,972	0,384	0,896	0,441	0,784	0,432	0,648	0,375	0,500	0,288	0,352	0,189	0,216	0,096	0,104	0,027	0,028	0,007	0,007
3	2	0,007	1,000	0,027	0,999	0,096	0,992	0,189	0,973	0,288	0,936	0,375	0,875	0,432	0,784	0,441	0,657	0,384	0,488	0,243	0,271	0,135	0,143
3	3	0,000	1,000	0,001	1,000	0,008	1,000	0,027	1,000	0,064	1,000	0,125	1,000	0,216	1,000	0,343	1,000	0,512	1,000	0,729	1,000	0,857	1,000
4	0	0,815	0,815	0,656	0,656	0,410	0,410	0,240	0,240	0,130	0,130	0,063	0,063	0,026	0,026	0,008	0,008	0,002	0,002	0,000	0,000	0,000	0,000
4	1	0,171	0,986	0,292	0,948	0,410	0,819	0,412	0,652	0,346	0,475	0,250	0,313	0,154	0,179	0,076	0,084	0,026	0,027	0,004	0,004	0,000	0,000
4	2	0,014	1,000	0,049	0,996	0,154	0,973	0,265	0,916	0,346	0,821	0,375	0,688	0,346	0,525	0,265	0,348	0,154	0,181	0,049	0,052	0,014	0,014
4	3	0,000	1,000	0,004	1,000	0,026	0,998	0,076	0,992	0,154	0,974	0,250	0,938	0,346	0,870	0,412	0,760	0,410	0,590	0,292	0,344	0,171	0,185
4	4	0,000	1,000	0,000	1,000	0,002	1,000	0,008	1,000	0,026	1,000	0,063	1,000	0,130	1,000	0,240	1,000	0,410	1,000	0,656	1,000	0,815	1,000

Продолжение табл. В.3

		$\pi = 0{,}05$		$\pi = 0{,}10$		$\pi = 0{,}20$		$\pi = 0{,}30$		$\pi = 0{,}40$		$\pi = 0{,}50$		$\pi = 0{,}60$		$\pi = 0{,}70$		$\pi = 0{,}80$		$\pi = 0{,}90$		$\pi = 0{,}95$	
n	*a*	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
5	0	0,774	0,774	0,590	0,590	0,328	0,328	0,168	0,168	0,078	0,078	0,031	0,031	0,010	0,010	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000
5	1	0,204	0,977	0,328	0,919	0,410	0,737	0,360	0,528	0,259	0,337	0,156	0,188	0,077	0,087	0,028	0,031	0,006	0,007	0,000	0,000	0,000	0,000
5	2	0,021	0,999	0,073	0,991	0,205	0,942	0,309	0,837	0,346	0,683	0,313	0,500	0,230	0,317	0,132	0,163	0,051	0,058	0,008	0,009	0,001	0,001
5	3	0,001	1,000	0,008	1,000	0,051	0,993	0,132	0,969	0,230	0,913	0,313	0,813	0,346	0,663	0,309	0,472	0,205	0,263	0,073	0,081	0,021	0,023
5	4	0,000	1,000	0,000	1,000	0,006	1,000	0,028	0,998	0,077	0,990	0,156	0,969	0,259	0,922	0,360	0,832	0,410	0,672	0,328	0,410	0,204	0,226
5	5	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,010	1,000	0,031	1,000	0,078	1,000	0,168	1,000	0,328	1,000	0,590	1,000	0,774	1,000
6	0	0,735	0,735	0,531	0,531	0,262	0,262	0,118	0,118	0,047	0,047	0,016	0,016	0,004	0,004	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
6	1	0,232	0,967	0,354	0,886	0,393	0,655	0,303	0,420	0,187	0,233	0,094	0,109	0,037	0,041	0,010	0,011	0,002	0,002	0,000	0,000	0,000	0,000
6	2	0,031	0,998	0,998	0,984	0,246	0,901	0,324	0,744	0,311	0,544	0,234	0,344	0,138	0,179	0,060	0,070	0,015	0,017	0,001	0,001	0,000	0,000
6	3	0,002	1,000	0,015	0,999	0,082	0,983	0,185	0,930	0,276	0,821	0,313	0,656	0,276	0,456	0,185	0,256	0,082	0,099	0,015	0,016	0,002	0,002
6	4	0,000	1,000	0,001	1,000	0,015	0,998	0,060	0,989	0,138	0,959	0,234	0,891	0,311	0,767	0,324	0,580	0,246	0,345	0,098	0,114	0,031	0,033
6	5	0,000	1,000	0,000	1,000	0,002	1,000	0,010	0,999	0,037	0,996	0,094	0,984	0,187	0,953	0,303	0,882	0,393	0,738	0,354	0,469	0,232	0,265
6	6	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,004	1,000	0,016	1,000	0,047	1,000	0,118	1,000	0,262	1,000	0,531	1,000	0,735	1,000
7	0	0,698	0,698	0,478	0,478	0,210	0,210	0,082	0,082	0,028	0,028	0,008	0,008	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
7	1	0,257	0,956	0,372	0,850	0,367	0,577	0,247	0,329	0,131	0,159	0,055	0,063	0,017	0,019	0,004	0,004	0,000	0,000	0,000	0,000	0,000	0,000
7	2	0,041	0,996	0,124	0,974	0,275	0,852	0,318	0,647	0,261	0,420	0,164	0,227	0,077	0,096	0,025	0,029	0,04	0,005	0,000	0,000	0,000	0,000
7	3	0,004	1,000	0,023	0,997	0,115	0,967	0,227	0,874	0,290	0,710	0,273	0,500	0,194	0,290	0,097	0,126	0,029	0,033	0,003	0,003	0,000	0,000
7	4	0,000	1,000	0,003	1,000	0,029	0,995	0,097	0,971	0,194	0,904	0,273	0,773	0,290	0,580	0,227	0,353	0,115	0,148	0,023	0,026	0,004	0,004
7	5	0,000	1,000	0,000	1,000	0,004	1,000	0,025	0,996	0,077	0,981	0,164	0,938	0,261	0,841	0,318	0,671	0,275	0,423	0,124	0,150	0,041	0,044

Продолжение табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
7	6	0,000	1,000	0,000	1,000	0,000	1,000	0,004	1,000	0,017	0,998	0,055	0,992	0,131	0,972	0,247	0,918	0,367	0,790	0,372	0,522	0,257	0,302
7	7	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,008	1,000	0,028	1,000	0,082	1,000	0,210	1,000	0,478	1,000	0,698	1,000
8	0	0,663	0,663	0,430	0,430	0,168	0,168	0,058	0,058	0,017	0,017	0,004	0,004	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
8	1	0,279	0,943	0,383	0,813	0,336	0,503	0,198	0,255	0,090	0,106	0,031	0,035	0,008	0,009	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
8	2	0,051	0,994	0,149	0,962	0,294	0,797	0,296	0,552	0,209	0,315	0,109	0,145	0,041	0,050	0,010	0,011	0,001	0,001	0,000	0,000	0,000	0,000
8	3	0,005	1,000	0,033	0,995	0,147	0,944	0,254	0,806	0,279	0,594	0,219	0,363	0,124	0,174	0,047	0,058	0,009	0,010	0,000	0,000	0,000	0,000
8	4	0,000	1,000	0,005	1,000	0,046	0,990	0,136	0,942	0,232	0,826	0,273	0,637	0,232	0,406	0,136	0,194	0,046	0,056	0,005	0,005	0,000	0,000
8	5	0,000	1,000	0,000	1,000	0,009	0,999	0,047	0,989	0,124	0,950	0,219	0,855	0,279	0,685	0,254	0,448	0,147	0,203	0,033	0,038	0,005	0,006
8	6	0,000	1,000	0,000	1,000	0,001	1,000	0,010	0,999	0,041	0,991	0,109	0,965	0,209	0,894	0,296	0,745	0,294	0,497	0,149	0,187	0,051	0,057
8	7	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,008	0,999	0,031	0,996	0,090	0,983	0,198	0,942	0,336	0,832	0,383	0,570	0,279	0,337
8	8	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,004	1,000	0,017	1,000	0,058	1,000	0,168	1,000	0,430	1,000	0,663	1,000
9	0	0,630	0,630	0,387	0,387	0,134	0,134	0,040	0,040	0,010	0,010	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
9	1	0,299	0,929	0,387	0,775	0,302	0,436	0,156	0,196	0,060	0,071	0,018	0,020	0,004	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
9	2	0,063	0,992	0,172	0,947	0,302	0,738	0,267	0,463	0,161	0,232	0,070	0,090	0,021	0,025	0,004	0,004	0,000	0,000	0,000	0,000	0,000	0,000
9	3	0,008	0,999	0,045	0,992	0,176	0,914	0,267	0,730	0,251	0,483	0,164	0,254	0,074	0,099	0,021	0,025	0,003	0,003	0,000	0,000	0,000	0,000
9	4	0,001	1,000	0,007	0,999	0,066	0,980	0,172	0,901	0,251	0,733	0,246	0,500	0,167	0,267	0,074	0,099	0,017	0,020	0,001	0,001	0,000	0,000
9	5	0,000	1,000	0,001	1,000	0,017	0,997	0,074	0,975	0,167	0,901	0,246	0,746	0,251	0,517	0,172	0,270	0,066	0,086	0,007	0,008	0,001	0,001
9	6	0,000	1,000	0,000	1,000	0,003	1,000	0,021	0,996	0,074	0,975	0,164	0,910	0,251	0,768	0,267	0,537	0,176	0,262	0,045	0,053	0,008	0,008
9	7	0,000	1,000	0,000	1,000	0,000	1,000	0,004	1,000	0,021	0,996	0,070	0,980	0,161	0,929	0,267	0,804	0,302	0,564	0,172	0,225	0,063	0,071

Продолжение табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
9	8	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,004	1,000	0,018	0,998	0,060	0,990	0,156	0,960	0,302	0,866	0,387	0,613	0,299	0,370
9	9	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,010	1,000	0,040	1,000	0,134	1,000	0,387	1,000	0,630	1,000
10	0	0,599	0,599	0,349	0,349	0,107	0,107	0,028	0,028	0,006	0,006	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10	1	0,315	0,914	0,387	0,736	0,268	0,376	0,121	0,149	0,040	0,046	0,010	0,011	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10	2	0,075	0,988	0,194	0,930	0,302	0,678	0,233	0,383	0,121	0,167	0,044	0,055	0,011	0,012	0,001	0,002	0,000	0,000	0,000	0,000	0,000	0,000
10	3	0,010	0,999	0,057	0,987	0,201	0,879	0,267	0,650	0,215	0,382	0,117	0,172	0,042	0,055	0,009	0,011	0,001	0,001	0,000	0,000	0,000	0,000
10	4	0,001	1,000	0,011	0,998	0,088	0,967	0,200	0,850	0,251	0,633	0,205	0,377	0,111	0,166	0,037	0,047	0,006	0,006	0,000	0,000	0,000	0,000
10	5	0,000	1,000	0,001	1,000	0,026	0,994	0,103	0,953	0,201	0,834	0,246	0,623	0,201	0,367	0,103	0,150	0,026	0,033	0,001	0,002	0,000	0,000
10	6	0,000	1,000	0,000	1,000	0,006	0,999	0,037	0,989	0,111	0,945	0,205	0,828	0,251	0,618	0,200	0,350	0,088	0,121	0,011	0,013	0,001	0,001
10	7	0,000	1,000	0,000	1,000	0,001	1,000	0,009	0,998	0,042	0,988	0,117	0,945	0,215	0,833	0,267	0,617	0,201	0,322	0,057	0,070	0,010	0,012
10	8	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,011	0,998	0,044	0,989	0,121	0,954	0,233	0,851	0,302	0,624	0,194	0,264	0,075	0,086
10	9	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,010	0,999	0,040	0,994	0,121	0,972	0,268	0,893	0,387	0,651	0,315	0,401
10	10	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,006	1,000	0,028	1,000	0,107	1,000	0,349	1,000	0,599	1,000
11	0	0,569	0,569	0,314	0,314	0,086	0,086	0,020	0,020	0,004	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
11	1	0,329	0,898	0,384	0,697	0,236	0,322	0,093	0,113	0,027	0,030	0,005	0,006	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
11	2	0,087	0,985	0,213	0,910	0,295	0,617	0,200	0,313	0,089	0,119	0,027	0,033	0,005	0,006	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
11	3	0,014	0,998	0,071	0,981	0,221	0,839	0,257	0,570	0,177	0,296	0,081	0,113	0,023	0,029	0,004	0,004	0,000	0,000	0,000	0,000	0,000	0,000
11	4	0,001	1,000	0,016	0,997	0,111	0,950	0,220	0,790	0,236	0,533	0,161	0,274	0,070	0,099	0,017	0,022	0,002	0,002	0,000	0,000	0,000	0,000
11	5	0,000	1,000	0,002	1,000	0,039	0,988	0,132	0,922	0,221	0,753	0,226	0,500	0,147	0,247	0,057	0,078	0,010	0,012	0,000	0,000	0,000	0,000

Продолжение табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
11	6	0,000	1,000	0,000	1,000	0,010	0,998	0,057	0,978	0,147	0,901	0,226	0,726	0,221	0,467	0,132	0,210	0,039	0,050	0,002	0,003	0,000	0,000
11	7	0,000	1,000	0,000	1,000	0,002	1,000	0,017	0,996	0,070	0,971	0,161	0,887	0,236	0,704	0,220	0,430	0,111	0,161	0,016	0,019	0,001	0,002
11	8	0,000	1,000	0,000	1,000	0,000	1,000	0,004	0,999	0,023	0,994	0,081	0,967	0,177	0,881	0,257	0,687	0,221	0,383	0,071	0,090	0,014	0,015
11	9	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,005	0,999	0,027	0,994	0,089	0,970	0,200	0,887	0,295	0,678	0,213	0,303	0,087	0,102
11	10	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,005	1,000	0,027	0,996	0,093	0,980	0,236	0,914	0,384	0,686	0,329	0,431
11	11	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,004	1,000	0,020	1,000	0,086	1,000	0,314	1,000	0,569	1,000
12	0	0,540	0,540	0,282	0,282	0,069	0,069	0,014	0,014	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
12	1	0,341	0,882	0,377	0,659	0,206	0,275	0,071	0,085	0,017	0,020	0,003	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
12	2	0,099	0,980	0,230	0,889	0,283	0,558	0,168	0,253	0,064	0,083	0,016	0,019	0,002	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
12	3	0,017	0,998	0,085	0,974	0,236	0,795	0,240	0,493	0,142	0,225	0,054	0,073	0,012	0,015	0,001	0,002	0,000	0,000	0,000	0,000	0,000	0,000
12	4	0,002	1,000	0,021	0,996	0,133	0,927	0,231	0,724	0,213	0,438	0,121	0,194	0,042	0,057	0,008	0,009	0,001	0,001	0,000	0,000	0,000	0,000
12	5	0,000	1,000	0,004	0,999	0,053	0,981	0,158	0,882	0,227	0,665	0,193	0,387	0,101	0,158	0,029	0,039	0,003	0,004	0,000	0,000	0,000	0,000
12	6	0,000	1,000	0,000	1,000	0,016	0,996	0,079	0,961	0,177	0,842	0,226	0,613	0,177	0,335	0,079	0,118	0,016	0,019	0,000	0,001	0,000	0,000
12	7	0,000	1,000	0,000	1,000	0,003	0,999	0,029	0,991	0,101	0,943	0,193	0,806	0,227	0,562	0,158	0,276	0,053	0,073	0,004	0,004	0,000	0,000
12	8	0,000	1,000	0,000	1,000	0,001	1,000	0,008	0,998	0,042	0,985	0,121	0,927	0,213	0,775	0,231	0,507	0,133	0,205	0,021	0,026	0,002	0,002
12	9	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,012	0,997	0,054	0,981	0,142	0,917	0,240	0,747	0,236	0,442	0,085	0,111	0,017	0,020
12	10	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,016	0,997	0,064	0,980	0,168	0,915	0,283	0,725	0,230	0,341	0,099	0,118
12	11	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,003	1,000	0,017	0,998	0,071	0,986	0,206	0,931	0,377	0,718	0,341	0,460
12	12	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,014	1,000	0,069	1,000	0,282	1,000	0,540	1,000

Продолжение табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
13	0	0,513	0,513	0,254	0,254	0,055	0,055	0,010	0,010	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
13	1	0,351	0,865	0,367	0,621	0,179	0,234	0,054	0,064	0,011	0,013	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
13	2	0,111	0,975	0,245	0,866	0,268	0,502	0,139	0,202	0,045	0,058	0,010	0,011	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
13	3	0,021	0,997	0,100	0,966	0,246	0,747	0,218	0,421	0,111	0,169	0,035	0,046	0,006	0,008	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
13	4	0,003	1,000	0,028	0,994	0,154	0,901	0,234	0,654	0,184	0,353	0,087	0,133	0,024	0,032	0,003	0,004	0,000	0,000	0,000	0,000	0,000	0,000
13	5	0,000	1,000	0,006	0,999	0,069	0,970	0,180	0,835	0,221	0,574	0,157	0,291	0,066	0,098	0,014	0,018	0,001	0,001	0,000	0,000	0,000	0,000
13	6	0,000	1,000	0,001	1,000	0,023	0,993	0,103	0,938	0,197	0,771	0,209	0,500	0,131	0,229	0,044	0,062	0,006	0,007	0,000	0,000	0,000	0,000
13	7	0,000	1,000	0,000	1,000	0,006	0,999	0,044	0,982	0,131	0,902	0,209	0,709	0,197	0,426	0,103	0,165	0,023	0,030	0,001	0,001	0,000	0,000
13	8	0,000	1,000	0,000	1,000	0,001	1,000	0,014	0,996	0,066	0,968	0,157	0,867	0,221	0,647	0,180	0,346	0,069	0,099	0,006	0,006	0,000	0,000
13	9	0,000	1,000	0,000	1,000	0,000	1,000	0,003	0,999	0,024	0,992	0,087	0,954	0,184	0,831	0,234	0,579	0,154	0,253	0,028	0,034	0,003	0,003
13	10	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,006	0,999	0,035	0,989	0,111	0,942	0,218	0,798	0,246	0,498	0,100	0,134	0,021	0,025
13	11	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,010	0,998	0,045	0,987	0,139	0,936	0,268	0,766	0,245	0,379	0,111	0,135
13	12	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,011	0,999	0,054	0,990	0,179	0,945	0,367	0,746	0,351	0,487
13	13	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,010	1,000	0,055	1,000	0,254	1,000	0,513	1,000
14	0	0,488	0,488	0,229	0,229	0,044	0,044	0,007	0,007	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
14	1	0,359	0,847	0,356	0,585	0,154	0,198	0,041	0,047	0,007	0,008	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
14	2	0,123	0,970	0,257	0,842	0,250	0,448	0,113	0,161	0,032	0,040	0,006	0,006	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
14	3	0,026	0,996	0,114	0,956	0,250	0,698	0,194	0,355	0,085	0,124	0,022	0,029	0,003	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
14	4	0,004	1,000	0,035	0,991	0,172	0,870	0,229	0,584	0,155	0,279	0,061	0,090	0,014	0,018	0,001	0,002	0,000	0,000	0,000	0,000	0,000	0,000
14	5	0,000	1,000	0,008	0,999	0,086	0,956	0,196	0,781	0,207	0,486	0,122	0,212	0,041	0,058	0,007	0,008	0,000	0,000	0,000	0,000	0,000	0,000

Продолжение табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
14	6	0,000	1,000	0,001	1,000	0,032	0,988	0,126	0,907	0,207	0,692	0,183	0,395	0,092	0,150	0,023	0,031	0,002	0,002	0,000	0,000	0,000	0,000
14	7	0,000	1,000	0,000	1,000	0,009	0,998	0,062	0,969	0,157	0,850	0,209	0,605	0,157	0,308	0,062	0,093	0,009	0,012	0,000	0,000	0,000	0,000
14	8	0,000	1,000	0,000	1,000	0,002	1,000	0,023	0,992	0,092	0,942	0,183	0,788	0,207	0,514	0,126	0,219	0,032	0,044	0,001	0,001	0,000	0,000
14	9	0,000	1,000	0,000	1,000	0,000	1,000	0,007	0,998	0,041	0,982	0,122	0,910	0,207	0,721	0,196	0,416	0,086	0,130	0,008	0,009	0,000	0,000
14	10	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,014	0,996	0,061	0,971	0,155	0,876	0,229	0,645	0,172	0,302	0,035	0,044	0,004	0,004
14	11	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,003	0,999	0,022	0,994	0,085	0,960	0,194	0,839	0,250	0,552	0,114	0,158	0,026	0,030
14	12	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,006	0,999	0,032	0,992	0,113	0,953	0,250	0,802	0,257	0,415	0,123	0,153
14	13	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,007	0,999	0,041	0,993	0,154	0,956	0,356	0,771	0,359	0,512
14	14	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,007	1,000	0,044	1,000	0,229	1,000	0,488	1,000
15	0	0,463	0,463	0,206	0,206	0,035	0,035	0,005	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	1	0,366	0,829	0,343	0,549	0,132	0,167	0,031	0,035	0,005	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	2	0,135	0,964	0,267	0,816	0,231	0,398	0,092	0,127	0,022	0,027	0,003	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	3	0,031	0,995	0,129	0,944	0,250	0,648	0,170	0,297	0,063	0,091	0,014	0,018	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	4	0,005	0,999	0,043	0,987	0,188	0,836	0,219	0,515	0,127	0,217	0,042	0,059	0,007	0,009	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
15	5	0,001	1,000	0,010	0,998	0,103	0,939	0,206	0,722	0,186	0,403	0,092	0,151	0,024	0,034	0,003	0,004	0,000	0,000	0,000	0,000	0,000	0,000
15	6	0,000	1,000	0,002	1,000	0,043	0,982	0,147	0,869	0,207	0,610	0,153	0,304	0,061	0,095	0,012	0,015	0,001	0,001	0,000	0,000	0,000	0,000
15	7	0,000	1,000	0,000	1,000	0,014	0,996	0,081	0,950	0,177	0,787	0,196	0,500	0,118	0,213	0,035	0,050	0,003	0,004	0,000	0,000	0,000	0,000
15	8	0,000	1,000	0,000	1,000	0,003	0,999	0,035	0,985	0,118	0,905	0,196	0,696	0,177	0,390	0,081	0,131	0,014	0,018	0,000	0,000	0,000	0,000
15	9	0,000	1,000	0,000	1,000	0,001	1,000	0,012	0,996	0,061	0,966	0,153	0,849	0,207	0,597	0,147	0,278	0,043	0,061	0,002	0,002	0,000	0,000
15	10	0,000	1,000	0,000	1,000	0,000	1,000	0,003	0,999	0,024	0,991	0,092	0,941	0,186	0,783	0,206	0,485	0,103	0,164	0,010	0,013	0,001	0,001

Продолжение табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
15	11	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,007	0,998	0,042	0,982	0,127	0,909	0,219	0,703	0,188	0,352	0,043	0,056	0,005	0,005
15	12	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,014	0,996	0,063	0,973	0,170	0,873	0,250	0,602	0,129	0,184	0,031	0,036
15	13	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,003	1,000	0,022	0,995	0,092	0,965	0,231	0,833	0,267	0,451	0,135	0,171
15	14	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,005	1,000	0,031	0,995	0,132	0,965	0,343	0,794	0,366	0,537
15	15	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,005	1,000	0,035	1,000	0,206	1,000	0,463	1,000
16	0	0,440	0,440	0,185	0,185	0,028	0,028	0,003	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
16	1	0,371	0,811	0,329	0,515	0,113	0,141	0,023	0,026	0,003	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
16	2	0,146	0,957	0,275	0,789	0,211	0,352	0,073	0,099	0,015	0,018	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
16	3	0,036	0,993	0,142	0,932	0,245	0,598	0,146	0,245	0,047	0,065	0,009	0,011	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
16	4	0,006	0,999	0,051	0,983	0,200	0,798	0,204	0,450	0,101	0,167	0,028	0,038	0,004	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
16	5	0,001	1,000	0,014	0,997	0,120	0,918	0,210	0,660	0,162	0,329	0,067	0,105	0,014	0,019	0,001	0,002	0,000	0,000	0,000	0,000	0,000	0,000
16	6	0,000	1,000	0,003	0,999	0,055	0,973	0,165	0,825	0,198	0,527	0,122	0,227	0,039	0,058	0,006	0,007	0,000	0,000	0,000	0,000	0,000	0,000
16	7	0,000	1,000	0,000	1,000	0,020	0,993	0,101	0,926	0,189	0,716	0,175	0,402	0,084	0,142	0,019	0,026	0,001	0,001	0,000	0,000	0,000	0,000
16	8	0,000	1,000	0,000	1,000	0,006	0,999	0,049	0,974	0,142	0,858	0,196	0,598	0,142	0,284	0,049	0,074	0,006	0,007	0,000	0,000	0,000	0,000
16	9	0,000	1,000	0,000	1,000	0,001	1,000	0,019	0,993	0,084	0,942	0,175	0,773	0,189	0,473	0,101	0,175	0,020	0,027	0,000	0,001	0,000	0,000
16	10	0,000	1,000	0,000	1,000	0,000	1,000	0,006	0,998	0,039	0,981	0,122	0,895	0,198	0,671	0,165	0,340	0,055	0,082	0,003	0,003	0,000	0,000
16	11	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,014	0,995	0,067	0,962	0,162	0,833	0,210	0,550	0,120	0,202	0,014	0,017	0,001	0,001
16	12	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,004	0,999	0,028	0,989	0,101	0,935	0,204	0,754	0,200	0,402	0,051	0,068	0,006	0,007
16	13	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,009	0,998	0,047	0,982	0,146	0,901	0,246	0,648	0,142	0,211	0,036	0,043
16	14	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,015	0,997	0,073	0,974	0,211	0,859	0,275	0,485	0,146	0,189

Продолжение табл. В.3

		π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
n	*a*	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
16	15	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,003	1,000	0,023	0,997	0,113	0,972	0,329	0,815	0,371	0,560
16	16	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,003	1,000	0,028	1,000	0,185	1,000	0,440	1,000
17	0	0,418	0,418	0,167	0,167	0,023	0,023	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
17	1	0,371	0,811	0,329	0,515	0,113	0,141	0,023	0,026	0,003	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
17	2	0,146	0,957	0,275	0,789	0,211	0,352	0,073	0,099	0,015	0,018	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
17	3	0,036	0,993	0,142	0,932	0,246	0,598	0,146	0,246	0,047	0,065	0,009	0,011	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
17	4	0,008	0,999	0,060	0,978	0,209	0,758	0,187	0,389	0,080	0,126	0,018	0,025	0,002	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
17	5	0,001	1,000	0,017	0,995	0,136	0,894	0,208	0,597	0,138	0,264	0,047	0,072	0,008	0,011	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
17	6	0,000	1,000	0,004	0,999	0,068	0,962	0,178	0,775	0,184	0,448	0,094	0,166	0,024	0,035	0,003	0,003	0,000	0,000	0,000	0,000	0,000	0,000
17	7	0,000	1,000	0,001	1,000	0,027	0,989	0,120	0,895	0,193	0,641	0,148	0,315	0,057	0,092	0,009	0,013	0,000	0,000	0,000	0,000	0,000	0,000
17	8	0,000	1,000	0,000	1,000	0,008	0,997	0,064	0,960	0,161	0,801	0,185	0,500	0,107	0,199	0,028	0,040	0,002	0,003	0,000	0,000	0,000	0,000
17	9	0,000	1,000	0,000	1,000	0,002	1,000	0,028	0,987	0,107	0,908	0,185	0,685	0,161	0,359	0,064	0,105	0,008	0,011	0,000	0,000	0,000	0,000
17	10	0,000	1,000	0,000	1,000	0,000	1,000	0,009	0,997	0,057	0,965	0,148	0,834	0,193	0,552	0,120	0,225	0,027	0,038	0,001	0,001	0,000	0,000
17	11	0,000	1,000	0,000	1,000	0,000	1,000	0,003	1,999	0,024	0,989	0,094	0,928	0,184	0,736	0,178	0,403	0,068	0,106	0,004	0,005	0,000	0,000
17	12	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,008	0,997	0,047	0,975	0,138	0,874	0,208	0,611	0,136	0,242	0,017	0,022	0,001	0,001
17	13	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,018	0,994	0,080	0,954	0,187	0,798	0,209	0,451	0,060	0,083	0,008	0,009
17	14	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,005	0,999	0,034	0,988	0,125	0,923	0,239	0,690	0,156	0,238	0,041	0,050
17	15	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,010	0,998	0,058	0,981	0,191	0,882	0,280	0,518	0,158	0,208
17	16	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,017	0,998	0,096	0,977	0,315	0,833	0,374	0,582
17	17	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,023	1,000	0,167	1,000	0,418	1,000

Продолжение табл. В.3

		π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
n	*a*	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
18	0	0,397	0,397	0,150	0,150	0,018	0,018	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
18	1	0,376	0,774	0,300	0,450	0,081	0,099	0,013	0,014	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
18	2	0,168	0,942	0,284	0,734	0,172	0,271	0,046	0,060	0,007	0,008	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
18	3	0,047	0,989	0,168	0,902	0,230	0,501	0,105	0,165	0,025	0,033	0,003	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
18	4	0,009	0,998	0,070	0,972	0,215	0,716	0,168	0,333	0,061	0,094	0,012	0,015	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
18	5	0,001	1,000	0,022	0,994	0,151	0,867	0,202	0,534	0,115	0,209	0,033	0,048	0,004	0,006	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
18	6	0,000	1,000	0,005	0,999	0,082	0,949	0,187	0,722	0,166	0,374	0,071	0,119	0,015	0,020	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
18	7	0,000	1,000	0,001	1,000	0,035	0,984	0,138	0,859	0,189	0,563	0,121	0,240	0,037	0,058	0,005	0,006	0,000	0,000	0,000	0,000	0,000	0,000
18	8	0,000	1,000	0,000	1,000	0,012	0,996	0,081	0,940	0,173	0,737	0,167	0,407	0,077	0,135	0,015	0,021	0,001	0,001	0,000	0,000	0,000	0,000
18	9	0,000	1,000	0,000	1,000	0,003	0,999	0,039	0,979	0,128	0,865	0,185	0,593	0,128	0,263	0,039	0,060	0,003	0,004	0,000	0,000	0,000	0,000
18	10	0,000	1,000	0,000	1,000	0,001	1,000	0,015	0,994	0,077	0,942	0,167	0,760	0,173	0,437	0,081	0,141	0,012	0,016	0,000	0,000	0,000	0,000
18	11	0,000	1,000	0,000	1,000	0,000	1,000	0,005	0,999	0,037	0,980	0,121	0,881	0,189	0,626	0,138	0,278	0,035	0,051	0,001	0,001	0,000	0,000
18	12	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,015	0,994	0,071	0,952	0,166	0,791	0,187	0,466	0,082	0,133	0,005	0,006	0,000	0,000
18	13	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,004	0,999	0,033	0,985	0,115	0,906	0,202	0,667	0,151	0,284	0,022	0,028	0,001	0,002
18	14	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,012	0,996	0,061	0,967	0,168	0,835	0,215	0,499	0,070	0,098	0,009	0,011
18	15	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,003	0,999	0,025	0,992	0,105	0,940	0,230	0,729	0,168	0,266	0,047	0,058
18	16	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,007	0,999	0,046	0,986	0,172	0,901	0,284	0,550	0,168	0,226
18	17	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,013	0,998	0,081	0,982	0,300	0,850	0,376	0,603
18	18	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,018	1,000	0.150	1,000	0,397	1,000
19	0	0,377	0,377	0,135	0,135	0,014	0,014	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Продолжение табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
19	1	0,377	0,755	0,285	0,420	0,068	0,083	0,009	0,010	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
19	2	0,179	0,933	0,285	0,705	0,154	0,237	0,036	0,046	0,005	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
19	3	0,053	0,987	0,180	0,885	0,218	0,455	0,087	0,133	0,017	0,023	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
19	4	0,011	0,998	0,080	0,965	0,218	0,673	0,149	0,282	0,047	0,070	0,007	0,010	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
19	5	0,002	1,000	0,027	0,991	0,164	0,837	0,192	0,474	0,093	0,163	0,022	0,032	0,002	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
19	6	0,000	1,000	0,007	0,998	0,095	0,932	0,192	0,666	0,145	0,308	0,052	0,084	0,008	0,012	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
19	7	0,000	1,000	0,001	1,000	0,044	0,977	0,153	0,818	0,180	0,488	0,096	0,180	0,024	0,035	0,002	0,003	0,000	0,000	0,000	0,000	0,000	0,000
19	8	0,000	1,000	0,000	1,000	0,017	0,993	0,098	0,916	0,180	0,667	0,144	0,324	0,053	0,088	0,008	0,011	0,000	0,000	0,000	0,000	0,000	0,000
19	9	0,000	1,000	0,000	1,000	0,005	0,998	0,051	0,967	0,146	0,814	0,176	0,500	0,098	0,186	0,022	0,033	0,001	0,002	0,000	0,000	0,000	0,000
19	10	0,000	1,000	0,000	1,000	0,001	1,000	0,022	0,989	0,098	0,912	0,176	0,676	0,146	0,333	0,051	0,084	0,005	0,007	0,000	0,000	0,000	0,000
19	11	0,000	1,000	0,000	1,000	0,000	1,000	0,008	0,997	0,053	0,965	0,144	0,820	0,180	0,512	0,098	0,182	0,017	0,023	0,000	0,000	0,000	0,000
19	12	0,000	1,000	0,000	1,000	0,000	1,000	0,002	0,999	0,024	0,988	0,096	0,916	0,180	0,692	0,153	0,334	0,044	0,068	0,001	0,002	0,000	0,000
19	13	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,008	0,997	0,052	0,968	0,145	0,837	0,192	0,526	0,095	0,163	0,007	0,009	0,000	0,000
19	14	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	0,999	0,022	0,990	0,093	0,930	0,192	0,718	0,164	0,327	0,027	0,035	0,002	0,002
19	15	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,007	0,998	0,047	0,977	0,149	0,867	0,218	0,545	0,080	0,115	0,011	0,013
19	16	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,002	1,000	0,017	0,995	0,087	0,954	0,218	0,763	0,180	0,295	0,053	0,067
19	17	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,005	0,999	0,036	0,990	0,154	0,917	0,285	0,580	0,179	0,245
19	18	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,009	0,999	0,068	0,986	0,285	0,865	0,377	0,623
19	19	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,014	1,000	0,135	1,000	0,377	1,000
20	0	0,358	0,358	0,122	0,122	0,012	0,012	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Окончание табл. В.3

n	a	π = 0,05		π = 0,10		π = 0,20		π = 0,30		π = 0,40		π = 0,50		π = 0,60		π = 0,70		π = 0,80		π = 0,90		π = 0,95	
		Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма	Точное	Сумма
20	1	0,377	0,736	0,270	0,392	0,058	0,069	0,007	0,008	0,000	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	2	0,189	0,925	0,285	0,677	0,137	0,206	0,028	0,035	0,003	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	3	0,060	0,984	0,190	0,867	0,205	0,411	0,072	0,107	0,012	0,016	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	4	0,013	0,997	0,090	0,957	0,218	0,630	0,130	0,238	0,035	0,051	0,005	0,006	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	5	0,002	1,000	0,032	0,989	0,175	0,804	0,179	0,416	0,075	0,126	0,015	0,021	0,001	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	6	0,000	1,000	0,009	0,998	0,109	0,913	0,192	0,608	0,124	0,250	0,037	0,058	0,005	0,006	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	7	0,000	1,000	0,002	1,000	0,055	0,968	0,164	0,772	0,166	0,416	0,074	0,132	0,015	0,021	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000
20	8	0,000	1,000	0,000	1,000	0,022	0,990	0,114	0,887	0,180	0,596	0,120	0,252	0,035	0,057	0,004	0,005	0,000	0,000	0,000	0,000	0,000	0,000
20	9	0,000	1,000	0,000	1,000	0,007	0,997	0,065	0,952	0,160	0,755	0,160	0,412	0,071	0,128	0,012	0,017	0,000	0,001	0,000	0,000	0,000	0,000
20	10	0,000	1,000	0,000	1,000	0,002	0,999	0,031	0,983	0,117	0,872	0,176	0,588	0,117	0,245	0,031	0,048	0,002	0,003	0,000	0,000	0,000	0,000
20	11	0,000	1,000	0,000	1,000	0,000	1,000	0,012	0,995	0,071	0,943	0,160	0,748	0,160	0,404	0,065	0,113	0,007	0,010	0,000	0,000	0,000	0,000
20	12	0,000	1,000	0,000	1,000	0,000	1,000	0,004	0,999	0,035	0,979	0,120	0,868	0,180	0,584	0,114	0,228	0,022	0,032	0,000	0,000	0,000	0,000
20	13	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,015	0,994	0,074	0,942	0,166	0,750	0,164	0,392	0,055	0,087	0,002	0,002	0,000	0,000
20	14	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,005	0,998	0,037	0,979	0,124	0,874	0,192	0,584	0,109	0,196	0,009	0,011	0,000	0,000
20	15	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,015	0,994	0,075	0,949	0,179	0,762	0,175	0,370	0,032	0,043	0,002	0,003
20	16	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,005	0,999	0,035	0,984	0,130	0,893	0,218	0,589	0,090	0,133	0,013	0,016
20	17	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,012	0,996	0,072	0,965	0,205	0,794	0,190	0,323	0,060	0,075
20	18	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,003	0,999	0,028	0,992	0,137	0,931	0,285	0,608	0,189	0,264
20	19	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,007	0,999	0,058	0,988	0,270	0,878	0,377	0,642
20	20	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,001	1,000	0,012	1,000	0,122	1,000	0,358	1,000

Таблица В.4. t-таблица

Доверительный интервал							
Двухсторонний	80%	90%	95%	98%	99%	99,8%	99,9%
Односторонний	90%	95%	97,5%	99%	99,5%	99,9%	99,95%
Уровень значимости проверки гипотезы							
Двухсторонний тест	0,20	0,10	0,05	0,02	0,01	0,002	0,001
Односторонний тест	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
Для одной выборки: n	В целом: степени свободы	Критические значения					
2	1	3,078	6,314	12,706	31,821	63,657	318,309	636,619
3	2	1,886	2,920	4,303	6,965	9,925	22,327	31,599
4	3	1,638	2,353	3,182	4,541	5,841	10,215	12,924
5	4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
6	5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
7	6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
8	7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
9	8	1,397	1,860	2,306	2,896	3,355	4,505	5,041
10	9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
11	10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
12	11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
13	12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
14	13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
15	14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
16	15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
17	16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
18	17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
19	18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
20	19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
21	20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
22	21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
23	22	1,321	1,717	2,074	2,508	2,819	3,505	3,792

Доверительный интервал							
Двухсторонний	80%	90%	95%	98%	99%	99,8%	99,9%
Односторонний	90%	95%	97,5%	99%	99,5%	99,9%	99,95%
Уровень значимости проверки гипотезы							
Двухсторонний тест	0,20	0,10	0,05	0,02	0,01	0,002	0,001
Односторонний тест	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
Для одной выборки: n	В целом: степени свободы	Критические значения					
24	23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
25	24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
26	25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
27	26	1,315	1,706	2,056	2,479	2,779	3,434	3,707
28	27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
29	28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
30	29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
31	30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	31	1,309	1,696	2,040	2,453	2,744	3,375	3,633
33	32	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	33	1,308	1,692	2,035	2,445	2,733	3,356	3,611
35	34	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	35	1,306	1,690	2,030	2,438	2,724	3,340	3,591
37	36	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	37	1,305	1,687	2,026	2,431	2,715	3,326	3,574
39	38	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	39	1,304	1,685	1,023	2,426	2,708	3,313	3,558
Бесконечность		1,282	1,645	1,960	2,326	2,576	3,090	3,291

Таблица В.5. R^2 **-таблица: критические значения для уровня значимости 5% (значимый результат)**

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	0,994									
4	0,902	0,997								
5	0,771	0,950	0,998							
6	0,658	0,864	0,966	0,999						
7	0,569	0,776	0,903	0,975	0,999					
8	0,499	0,698	0,832	0,924	0,980	0,999				
9	0,444	0,632	0,764	0,865	0,938	0,983	0,999			
10	0,399	0,575	0,704	0,806	0,887	0,947	0,985	0,999		
11	0,362	0,527	0, 651	0,751	0,835	0,902	0, 954	0,987	1,000	
12	0,332	0,486	0, 604	0,702	0,785	0,856	0, 914	0,959	0,989	1,000
13	0,306	0,451	0,563	0,657	0,739	0,811	0,872	0,924	0,964	0,990
14	0,283	0,420	0,527	0,618	0,697	0,768	0,831	0,885	0,931	0,967
15	0,264	0,393	0,495	0,582	0,659	0,729	0,791	0,847	0,896	0937
16	0,247	0,369	0,466	0,550	0,624	0,692	0,754	0,810	0,860	0,904
17	0,232	0,348	0,440	0,521	0,593	0,659	0,719	0,775	0,825	0,871
18	0,219	0,329	0,417	0,494	0,564	0,628	0,687	0,742	0,792	0,839
19	0,208	0,312	0,397	0,471	0,538	0,600	0,657	0,711	0,761	0,807
20	0,197	0,297	0,378	0,449	0,514	0,574	0,630	0,682	0,731	0,777
21	0,187	0,283	0,361	0,429	0,492	0,550	0,604	0,655	0,703	0,749
22	0,179	0,270	0,345	0,411	0,471	0,527	0,580	0,630	0,677	0,722
23	0,171	0,259	0,331	0,394	0,452	0,507	0,558	0,607	0,653	0,696
24	0,164	0,248	0,317	0,379	0,435	0,488	0,538	0,585	0,630	0,673
25	0,157	0,238	0,305	0,364	0,419	0,470	0,518	0,564	0,608	0,650
26	0,151	0,229	0,294	0,351	0,404	0,454	0,501	0,545	0,588	0,629
27	0,145	0,221	0,283	0,339	0,390	0,438	0,484	0,527	0,569	0,609
28	0,140	0,213	0,273	0,327	0,377	0,424	0,468	0,510	0,551	0,590
29	0,135	0,206	0,264	0,316	0,365	0,410	0,453	0,495	0,534	0,573
30	0,130	0,199	0,256	0,306	0,353	0,397	0,439	0,480	0,518	0,556

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
31	0,126	0,193	0,248	0,297	0,342	0,385	0,426	0,466	0,503	0,540
32	0,122	0,187	0,240	0,288	0,332	0,374	0,414	0,452	0,489	0,525
33	0,118	0,181	0,233	0,279	0,323	0,363	0,402	0,440	0,476	0,511
34	0,115	0,176	0,226	0,271	0,314	0,353	0,391	0,428	0,463	0,497
35	0,111	0,171	0,220	0,264	0,305	0,344	0,381	0,417	0,451	0,484
36	0,108	0,166	0,214	0,257	0,297	0,335	0,371	0,406	0,440	0,472
37	0,105	0,162	0,208	0,250	0,289	0,326	0,362	0,396	0,429	0,461
38	0,103	0,157	0,203	0,244	0,282	0,318	0,353	0,386	0,418	0,449
39	0,100	0,153	0,198	0,238	0,275	0,310	0,344	0,377	0,408	0,439
40	0,097	0,150	0,193	0,232	0,268	0,303	0,336	0,368	0,399	0,429
41	0,095	0,146	0,188	0,226	0,262	0,296	0,328	0,359	0,390	0,419
42	0,093	0,142	0,184	0,221	0,256	0,289	0,321	0,351	0,381	0,410
43	0,090	0,139	0,180	0,216	0,250	0,283	0,314	0,344	0,373	0,401
44	0,088	0,136	0,176	0,211	0,245	0,276	0,307	0,336	0,365	0,393
45	0,086	0,133	0,172	0,207	0,239	0,271	0,300	0,329	0,357	0,384
46	0,085	0,130	0,168	0,202	0,234	0,265	0,294	0,322	0,350	0,377
47	0,083	0,127	0,164	0,198	0,230	0,259	0,288	0,316	0,343	0,369
48	0,081	0,125	0,161	0,194	0,225	0,254	0,282	0,310	0,336	0,362
49	0,079	0,122	0,158	0,190	0,220	0,249	0,277	0,304	0,330	0,355
50	0,078	0,120	0,155	0,186	0,216	0,244	0,272	0,298	0,323	0,348
51	0,076	0,117	0,152	0,183	0,212	0,240	0,267	0,293	0,318	0,342
52	0,075	0,115	0,149	0,180	0,208	0,235	0,262	0,287	0,312	0,336
53	0,073	0,113	0,146	0,176	0,204	0,231	0,257	0,282	0,306	0,330
54	0,072	0,111	0,143	0,173	0,201	0,227	0,252	0,277	0,301	0,324
55	0,071	0,109	0,141	0,170	0,197	0,223	0,248	0,272	0,295	0,318
56	0,069	0,107	0,138	0,167	0,194	0,219	0,244	0,267	0,290	0,313
57	0,068	0,105	0,136	0,164	0,190	0,215	0,240	0,263	0,285	0,308
58	0,067	0,103	0,134	0,161	0,187	0,212	0,236	0,258	0,281	0,303

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
59	0,066	0,101	0,131	0,159	0,184	0,208	0,232	0,254	0,276	0,298
60	0,065	0,100	0,129	0,156	0,181	0,205	0,228	0,250	0,272	0,293
Множитель										
1	3,84	5,99	7,82	9,49	11,07	12,59	14,07	15,51	16,92	18,31
2	2,15	−0,27	−3,84	−7,94	−12,84	−18,24	−23,78	−30,10	−36,87	−43,87

Таблица B.6. R^2-таблица: критические значения для уровня значимости 1% (высоко значимый результат)

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	1,000									
4	0,980	1,000								
5	0,919	0,990	1,000							
6	0,841	0,954	0,993	1,000						
7	0,765	0,900	0,967	0,995	1,000					
8	0,696	0,842	0,926	0,975	0,996	1,000				
9	0,636	0,785	0,879	0,941	0,979	0,997	1,000			
10	0,585	0,732	0,830	0,901	0,951	0,982	0,997	1,000		
11	0,540	0,684	0,784	0,859	0,916	0,958	0,985	0,997	1,000	
12	0,501	0,641	0,740	0,818	0,879	0,928	0,963	0,987	0,998	1,000
13	0,467	0,602	0,700	0,778	0,842	0,894	0,936	0,067	0,988	0,998
14	0,437	0,567	0,663	0,741	0,806	0,860	0,906	0,943	0,971	0,989
15	0,411	0,536	0,629	0,706	0,771	0,827	0,875	0,915	0,948	0,973
16	0,388	0,508	0,598	0,673	0,738	0,795	0,844	0,887	0,923	0,953
17	0,367	0,482	0,570	0,643	0,707	0,764	0,814	0,858	0,896	0,929
18	0,348	0,459	0,544	0,616	0,678	0,734	0,784	0,829	0,869	0,904
19	0,331	0,438	0,520	0,590	0,652	0,707	0,757	0,802	0,843	0,879
20	0,315	0,418	0,498	0,566	0,626	0,681	0,730	0,775	0,816	0,854

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
21	0,301	0,401	0,478	0,544	0,603	0,656	0,705	0,750	0,791	0,829
22	0,288	0,384	0,459	0,523	0,581	0,633	0,681	0,726	0,767	0,805
23	0,276	0,369	0,442	0,504	0,560	0,612	0,659	0,703	0,744	0,782
24	0,265	0,355	0,426	0,487	0,541	0,591	0,638	0,681	0,721	0,759
25	0,255	0,342	0,410	0,470	0,523	0,572	0,618	0,660	0,700	0,738
26	0,246	0,330	0,396	0,454	0,506	0,554	0,599	0,641	0,680	0,717
27	0,237	0,319	0,383	0,440	0,490	0,537	0,581	0,622	0,661	0,698
28	0,229	0,308	0,371	0,426	0,475	0,521	0,564	0,605	0,643	0,679
29	0,221	0,298	0,359	0,413	0,461	0,506	0,548	0,588	0,625	0,661
30	0,214	0,289	0,349	0,401	0,448	0,492	0,533	0,572	0,609	0,644
31	0,208	0,280	0,338	0,389	0,435	0,478	0,519	0,557	0,593	0,627
32	0,201	0,272	0,329	0,378	0,423	0, 465	0,505	0,542	0,578	0,612
33	0,195	0,264	0,319	0,368	0,412	0, 453	0,492	0,529	0,563	0,597
34	0,190	0,257	0,311	0,358	0,401	0, 442	0,479	0,515	0,550	0,583
35	0,185	0,250	0,303	0,349	0,391	0, 430	0,468	0,503	0,537	0,569
36	0,180	0,244	0,295	0,340	0,381	0,420	0,456	0,491	0,524	0,556
37	0,175	0,237	0,287	0,332	0,372	0,410	0,446	0,480	0,512	0,543
38	0,170	0,231	0,280	0,324	0,363	0,400	0,435	0,469	0,501	0,531
39	0,166	0,226	0,274	0,316	0,355	0,391	0,426	0,458	0,490	0,520
40	0,162	0,220	0,267	0,309	0,347	0,382	0,416	0,448	0,479	0,509
41	0,158	0,215	0,261	0,302	0,339	0,374	0,407	0,439	0,469	0,498
42	0,155	0,210	0,255	0,295	0,332	0,366	0,399	0,430	0,459	0,488
43	0,151	0,206	0,250	0,289	0,325	0,358	0,390	0,421	0,450	0,478
44	0,148	0,201	0,244	0,283	0,318	0,351	0,382	0,412	0,441	0,469
45	0,145	0,197	0,239	0,277	0,311	0,344	0,375	0,404	0,432	0,460
46	0,141	0,193	0,234	0,271	0,305	0,337	0,367	0,396	0,424	0,451
47	0,138	0,189	0,230	0,266	0,299	0,330	0,360	0,389	0,416	0,443
48	0,136	0,185	0,225	0,261	0,293	0,324	0,353	0,381	0,408	0,435

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
49	0,133	0,181	0,221	0,256	0,288	0,318	0,347	0,374	0,401	0,427
50	0,130	0,178	0,217	0,251	0,283	0,312	0,341	0,368	0,394	0,419
51	0,128	0,175	0,213	0,246	0,278	0,307	0,335	0,361	0,387	0,412
52	0,125	0,171	0,209	0,242	0,273	0,301	0,329	0,355	0,381	0,405
53	0,123	0,168	0,205	0,238	0,268	0,296	0,323	0,349	0,374	0,398
54	0,121	0,165	0,201	0,233	0,263	0,291	0,318	0,343	0,368	0,391
55	0,119	0,162	0,198	0,229	0,259	0,286	0,312	0,337	0,362	0,385
56	0,117	0,160	0,194	0,226	0,254	0,281	0,307	0,332	0,356	0,379
57	0,115	0,157	0,191	0,222	0,250	0,277	0,302	0,326	0,350	0,373
58	0,113	0,154	0,188	0,218	0,246	0,272	0,297	0,321	0,345	0,367
59	0,111	0,152	0,185	0,215	0,242	0,268	0,293	0,316	0,339	0,361
60	0,109	0,149	0,182	0,211	0,238	0,264	0,288	0,311	0,334	0,356
Множитель										
1	6,63	9,21	11,35	13,28	15,09	16,81	18,48	20,09	21,67	23,21
2	–5,81	–15,49	–25,66	–36,39	–47,63	–59,53	–71,65	–84,60	–97,88	-111,76

Таблица B.7. R^2-таблица: критические значения для уровня значимости 0,1% (очень высоко значимый результат)

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	1,000									
4	0,998	1,000								
5	0,982	0,999	1,000							
6	0,949	0,990	0,999	1,000						
7	0,904	0,968	0,993	0,999	1,000					
8	0,855	0,937	0,977	0,995	1,000	1,000				
9	0,807	0,900	0,952	0,982	0,996	1,000	1,000			
10	0,761	0,861	0,922	0,961	0,985	0,996	1,000	1,000		

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
11	0,717	0,822	0,889	0,936	0,967	0,987	0,997	1,000	1,000	
12	0,678	0,785	0,856	0,908	0,945	0,972	0,989	0,997	1,000	1,000
13	0,642	0,749	0,822	0,878	0,920	0,952	0,975	0,990	0,997	1,000
14	0,608	0,715	0,790	0,848	0,894	0,930	0,958	0,978	0,991	0,998
15	0,578	0,684	0,759	0,819	0,867	0,906	0,938	0,962	0,980	0,992
16	0,550	0,654	0,730	0,790	0,840	0,881	0,916	0,944	0,966	0,982
17	0,525	0,627	0,702	0,763	0,813	0,856	0,893	0,923	0,949	0,968
18	0,502	0,602	0,676	0,736	0,787	0,831	0,869	0,902	0,930	0,953
19	0,480	0,578	0,651	0,711	0,763	0,807	0,846	0,880	0,910	0,935
20	0,461	0,556	0,628	0,688	0,739	0,784	0,824	0,859	0,890	0,917
21	0,442	0,536	0,606	0,665	0,716	0,761	0,801	0,837	0,869	0,897
22	0,426	0,517	0,586	0,644	0,694	0,739	0,780	0,816	0,849	0,878
23	0,410	0,499	0,567	0,624	0,674	0,718	0,759	0,795	0,829	0,859
24	0,395	0,482	0,548	0,605	0,654	0,698	0,739	0,775	0,809	0,839
25	0,382	0,466	0,531	0,587	0,635	0,679	0,719	0,756	0,790	0,821
26	0,369	0,452	0,515	0,570	0,618	0,661	0,701	0,737	0,771	0,802
27	0,357	0,438	0,500	0,553	0,601	0,644	0,683	0,719	0,753	0,784
28	0,346	0,425	0,486	0,538	0,585	0,627	0,666	0,702	0,735	0,767
29	0,335	0,412	0,472	0,523	0,569	0,611	0,649	0,685	0,718	0,750
30	0,325	0,401	0,459	0,510	0,555	0,596	0,634	0,669	0,702	0,733
31	0,316	0,389	0,447	0,496	0,541	0,581	0,619	0,654	0,686	0,717
32	0,307	0,379	0,435	0,484	0,527	0,567	0,604	0,639	0,671	0,702
33	0,299	0,369	0,424	0,472	0,515	0,554	0,590	0,625	0,657	0,687
34	0,291	0,360	0,414	0,460	0,503	0,541	0,577	0,611	0,643	0,673
35	0,283	0,351	0,404	0,450	0,491	0,529	0,564	0,598	0,629	0,659
36	0,275	0,342	0,394	0,439	0,480	0,517	0,552	0,585	0,616	0,646
37	0,269	0,334	0,385	0,429	0,469	0,506	0,540	0,573	0,604	0,633
38	0,263	0,326	0,376	0,420	0,459	0,495	0,529	0,561	0,591	0,620
39	0,257	0,319	0,368	0,411	0,449	0,485	0,518	0,550	0,580	0,608
40	0,251	0,312	0,360	0,402	0,440	0,475	0,508	0,539	0,569	0,597

Окончание табл. B.7

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
41	0,245	0,305	0,352	0,393	0,431	0,465	0,498	0,529	0,558	0,586
42	0,240	0,298	0,345	0,385	0,422	0,456	0,488	0,518	0,547	0,575
43	0,235	0,292	0,338	0,378	0,414	0,447	0,479	0,509	0,537	0,564
44	0,230	0,286	0,331	0,370	0,406	0,439	0,470	0,499	0,527	0,554
45	0,225	0,280	0,324	0,363	0,398	0,431	0,461	0,490	0,518	0,544
46	0,220	0,275	0,318	0,356	0,391	0,423	0,453	0,482	0,509	0,535
47	0,216	0,269	0,312	0,349	0,383	0,415	0,445	0,473	0,500	0,526
48	0,212	0,264	0,306	0,343	0,377	0,408	0,437	0,465	0,491	0,517
49	0,208	0,259	0,301	0,337	0,370	0,401	0,429	0,457	0,483	0,508
50	0,204	0,255	0,295	0,331	0,363	0,394	0,422	0,449	0,475	0,500
51	0,200	0,250	0,290	0,325	0,357	0,387	0,415	0,442	0,467	0,492
52	0,197	0,246	0,285	0,320	0,351	0,381	0,408	0,435	0,460	0,484
53	0,193	0,242	0,280	0,314	0,345	0,374	0,402	0,428	0,453	0,477
54	0,190	0,237	0,276	0,309	0,340	0,368	0,395	0,421	0,446	0,469
55	0,186	0,233	0,271	0,304	0,334	0,362	0,389	0,414	0,439	0,462
56	0,183	0,230	0,267	0,299	0,329	0,357	0,383	0,408	0,432	0,455
57	0,180	0,226	0,262	0,294	0,324	0,351	0,377	0,402	0,426	0,448
58	0,177	0,222	0,258	0,290	0,319	0,346	0,371	0,396	0,419	0,442
59	0,174	0,219	0,254	0,285	0,314	0,341	0,366	0,390	0,413	0,436
60	0,172	0,215	0,250	0,281	0,309	0,336	0,361	0,384	0,407	0,429
Множитель										
1	10,83	13,82	16,27	18,47	20,52	22,46	24,32	26,12	27,88	29,59
2	−31,57	−54,02	−75,12	−96,26	−117,47	−138,94	−160,86	−183,33	−206,28	−229,55

Таблица B.8. R^2-таблица: критические значения для уровня значимости 10%

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
3	0,976									
4	0,810	0,990								
5	0,649	0,900	0,994							

Продолжение табл. B.8

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
6	0,532	0,785	0, 932	0,996						
7	0,448	0,684	0, 844	0,949	0,997					
8	0,386	0,602	0,759	0,877	0,959	0,997				
9	0,339	0,536	0,685	0,804	0,898	0,965	0,998			
10	0,302	0,482	0,622	0,738	0,835	0,914	0,970	0,998		
11	0,272	0,438	0,568	0,680	0,775	0,857	0,925	0, 974	0,998	
12	0,247	0,401	0,523	0,626	0,721	0,803	0,874	0, 933	0,977	0,998
13	0,227	0,369	0,484	0,584	0,673	0,753	0,825	0,888	0,940	0,979
14	0,209	0,342	0,450	0,545	0,630	0,708	0,779	0,842	0,899	0,946
15	0,194	0,319	0,420	0,510	0,592	0,667	0,736	0,799	0,857	0,907
16	0,181	0,298	0,394	0,480	0,558	0,630	0,697	0,759	0,816	0,868
17	0,170	0,280	0,371	0,453	0,527	0,596	0,661	0,721	0,778	0,830
18	0,160	0,264	0,351	0,428	0,499	0,566	0,628	0,687	0,742	0,794
19	0,151	0,250	0,332	0,406	0,474	0,538	0,598	0,655	0,709	0,760
20	0,143	0,237	0,316	0,386	0,452	0,513	0,571	0,626	0,679	0,729
21	0,136	0,226	0,301	0,368	0,431	0,490	0,546	0,599	0,650	0,699
22	0,129	0,215	0,287	0,352	0,412	0,469	0,523	0,575	0,624	0,671
23	0,124	0,206	0,275	0,337	0,395	0,450	0,502	0,552	0,600	0,646
24	0,118	0,197	0,263	0,323	0,379	0,432	0,486	0,530	0,577	0,622
25	0,113	0,189	0,253	0,310	0,364	0,415	0,464	0,511	0,556	0,599
26	0,109	0,181	0,243	0,298	0,350	0,400	0,447	0,492	0,536	0,579
27	0,105	0,175	0,234	0,287	0,338	0,386	0,431	0,475	0,518	0,559
28	0,101	0,168	0,225	0,277	0,326	0,372	0,417	0,459	0,501	0,541
29	0,097	0,162	0,218	0,268	0,315	0,360	0,403	0,444	0,484	0,523
30	0,094	0,157	0,210	0,259	0,305	0,348	0,390	0,430	0,469	0,507
31	0,091	0,152	0,203	0,251	0,295	0,337	0,378	0,417	0,455	0,492
32	0,088	0,147	0,197	0,243	0,286	0,327	0,366	0,405	0,442	0,478
33	0,085	0,142	0,191	0,236	0,277	0,317	0,356	0,393	0,429	0,464
34	0,082	0,138	0,185	0,229	0,269	0,308	0,346	0,382	0,417	0,451
35	0,080	0,134	0,180	0,222	0,262	0,300	0,336	0,371	0,406	0,439

Окончание табл. В.8

Количество наблюдений, n	Число X-переменных (k)									
	1	2	3	4	5	6	7	8	9	10
36	0,078	0,130	0,175	0, 216	0,255	0,291	0,327	0,361	0,395	0,427
37	0,075	0,127	0,170	0,210	0,248	0,284	0,318	0,352	0,385	0,416
38	0,073	0,123	0,166	0,205	0,241	0,276	0,310	0,343	0,375	0,406
39	0,071	0,120	0,162	0,199	0,235	0,269	0,302	0,334	0,366	0,396
40	0,070	0,117	0,157	0,194	0,229	0,263	0,295	0,326	0,357	0,387
41	0,068	0,114	0,154	0,190	0,224	0,257	0,288	0,319	0,348	0,378
42	0,066	0,111	0,150	0,185	0,219	0,250	0,281	0,311	0,340	0,369
43	0,065	0,109	0,146	0,181	0,214	0,245	0,275	0,304	0,333	0,361
44	0,063	0,106	0,143	0,177	0,209	0,239	0,269	0,297	0,325	0,353
45	0,062	0,104	0,140	0,173	0,204	0,234	0,263	0,291	0,318	0,345
46	0,0,060	0,102	0,137	0,169	0,200	0,229	0,257	0,285	0,312	0,338
47	0,059	0,099	0,134	0,166	0,196	0,224	0,252	0,279	0,305	0,331
48	0,058	0,097	0,131	0,162	0,191	0,220	0,247	0,273	0,299	0,324
49	0,057	0,095	0,128	0,159	0,188	0,215	0,242	0,268	0,293	0,318
50	0,055	0,093	0,126	0,156	0,184	0,211	0,237	0,263	0,287	0,312
51	0,054	0, 092	0,123	0,153	0,180	0,207	0,233	0,258	0,282	0,306
52	0,053	0,090	0,121	0,150	0,177	0,203	0,228	0,253	0,277	0,300
53	0,052	0,088	0,119	0,147	0,174	0,199	0,224	0,248	0,272	0,295
54	0,051	0,086	0,116	0,144	0,170	0,196	0,220	0,244	0,267	0,290
55	0,050	0,085	0,114	0,142	0,167	0,192	0,216	0,239	0,262	0,284
56	0,049	0,083	0,112	0,139	0,164	0,189	0,212	0,235	0,257	0,279
57	0,049	0,082	0,110	0,137	0,162	0,185	0,209	0,231	0,253	0,275
58	0,048	0,080	0,108	0,134	0,159	0,182	0,205	0,227	0,249	0,270
59	0,047	0,079	0,107	0,132	0,156	0,179	0,202	0,223	0,245	0,266
60	0,046	0,078	0,105	0,130	0,153	0,176	0,198	0,220	0,241	0,261
Множитель										
1	2,71	4,62	6,25	7,78	9,24	10,65	12,02	13,36	14,68	15,99
2	3,12	3,08	2,00	0,32	–1,92	–4,75	–7,59	–11,12	–14,94	–19,05

Таблица В.9. F-таблица: критические значения для уровня значимости 5% (значимый результат)

Знаменатель: степени свободы	Числитель: степени свободы																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	243,91	245,95	248,01	250,10	252,20	253,25	254,32
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396	19,413	19,429	19,446	19,462	19,479	19,487	19,496
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786	8,745	8,703	8,660	8,617	8,572	8,549	8,526
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,912	5,858	5,803	5,746	5,688	5,658	5,628
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,678	4,619	4,558	4,496	4,431	4,398	4,365
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,000	3,938	3,874	3,808	3,740	3,705	3,669
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,575	3,511	3,445	3,376	3,304	3,267	3,230
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347	3,284	3,218	3,150	3,079	3,005	2,967	2,928
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,073	3,006	2,936	2,864	2,787	2,748	2,707
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,913	2,845	2,774	2,700	2,621	2,680	2,538
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,687	2,617	2,544	2,466	2,384	2,341	2,296
15	4,543	3,682	3,287	3,056	2,901	2,780	2,707	2,641	2,588	2,544	2,475	2,403	2,328	2,247	2,160	2,114	2,066
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,278	2,203	2,124	2,039	1,946	1,896	1,843
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,092	2,015	1,932	1,841	1,740	1,683	1,622
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,917	1,836	1,748	1,649	1,534	1,467	1,389
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910	1,834	1,750	1,659	1,554	1,429	1,352	1,254
∞	3,841	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880	1,831	1,752	1,666	1,571	1,459	1,318	1,221	1,000

Таблица В.10. F-таблица: критические значения для уровня значимости 1% (высоко значимый результат)

Знаменатель: степени свободы	Числитель: степени свободы																
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
1	4052,2	4999,5	5403,4	5624,6	5763,7	5859,0	5928,4	5891,1	6022,5	6055,8	6106,3	6157,3	6208,7	6260,6	6313,0	6339,4	6365,9
2	98,501	98,995	99,159	99,240	99,299	99,333	99,356	99,374	99,388	99,399	99,416	99,432	99,449	99,466	99,482	99,491	99,499
3	34,116	30,816	29,456	28,709	28,236	27,910	27,671	27,488	27,344	27,228	27,051	26,871	26,689	26,503	26,315	26,220	26,125
4	21,197	18,00	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546	14,374	14,198	14,020	13,838	13,652	13,558	13,463
5	16,258	13,274	12,060	11,392	10,967	10,672	10,455	10,289	10,158	10,051	9,888	9,722	9,553	9,379	9,202	9,112	9,021
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,718	7,559	7,396	7,229	7,057	6,969	6,880
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,469	6,314	6,155	5,992	5,823	5,737	5,650
8	11,258	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,667	5,515	5,359	5,198	5,032	4,946	4,859
9	10,561	8,021	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,247	5,111	4,962	4,808	4,649	4,483	4,398	4,311
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,706	4,558	4,405	4,247	4,082	3,996	3,909
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	4,155	4,010	3,858	3,701	3,535	3,449	3,361
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,666	3,522	3,372	3,214	3,047	2,959	2,868
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	3,231	3,088	2,938	2,778	2,608	2,517	2,421
30	7,562	5,390	4,510	4,018	3,699	3,473	3,304	3,173	3,067	2,979	2,843	2,700	2,549	2,386	2,208	2,111	2,006
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632	2,496	2,352	2,198	2,028	1,836	1,726	1,601
120	6,851	4,786	3,949	3,480	3,174	2,956	2,792	2,663	2,559	2,472	2,336	2,191	2,035	1,860	1,656	1,533	1,381
∞	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407	2,321	2,185	2,039	1,878	1,696	1,473	1,325	1,000

Таблица В.11. F-таблица: критические значения для уровня значимости 0,1% (очень высоко значимый результат)

Знаменатель: степени свободы	Числитель: степени свободы																
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
1	405284	500000	540379	562500	576405	585937	592873	598144	602284	605621	610668	615764	620908	626099	631337	633972	636629
2	998,50	999,0	999,17	999,25	999,30	999,33	999,36	999,38	999,39	999,40	999,42	999,43	999,45	999,47	999,48	999,49	999,50
3	167,03	148,50	141,11	137,10	134,58	132,85	131,58	130,62	129,86	129,25	128,32	127,37	126,42	125,45	124,47	123,97	123,47
4	74,137	61,246	56,177	53,436	51,712	50,525	49,658	48,996	48,475	48,053	47,412	46,761	46,100	45,429	44,746	44,400	44,051
5	47,181	37,122	33,202	31,085	29,752	28,834	28,163	27,649	27,244	26,917	26,418	25,911	25,395	24,869	24,333	24,060	23,785
6	35,507	27,000	23,703	21,924	20,803	20,030	19,463	19,030	18,688	18,411	17,989	17,559	17,120	16,672	16,214	15,981	15,745
7	29,245	21,689	18,772	17,198	16,206	15,521	15,019	14,634	14,330	14,083	13,707	13,324	12,932	12,530	12,119	11,909	11,697
8	25,415	18,494	15,829	14,392	13,485	12,858	12,398	12,046	11,767	11,540	11,194	10,841	10,480	10,109	9,727	9,532	9,334
9	22,857	16,387	13,902	12,560	11,714	11,128	10,698	10,368	10,107	9,894	9,570	9,238	8,898	8,548	8,187	8,001	7,813
10	21,040	14,905	12,553	11,283	10,481	9,926	9,517	9,204	8,956	8,754	8,445	8,129	7,804	7,469	7,122	6,944	6,762
12	18,643	12,974	10,804	9,633	8,892	8,379	8,001	7,710	7,480	7,292	7,005	6,709	6,405	6,090	5,762	5,593	5,420
15	16,587	11,339	9,335	8,253	7,567	7,092	6,741	6,471	6,256	6,081	5,812	5,535	5,248	4,950	4,638	4,475	4,307
20	14,818	9,953	8,098	7,095	6,460	6,018	5,692	5,440	5,239	5,975	4,823	4,562	4,290	4,005	3,703	3,544	3,378
30	13,293	8,773	7,054	6,124	5,534	5,122	4,817	4,581	4,393	4,239	4,000	3,753	3,493	3,217	2,920	2,759	2,589
60	11,973	7,767	6,171	5,307	4,757	4,372	4,086	3,865	3,687	3,541	3,315	3,078	2,827	2,555	2,252	2,082	1,890
120	11,378	7,321	5,781	4,947	4,416	4,044	3,767	3,552	3,379	3,237	3,016	2,783	2,534	2,262	1,950	1,767	1,543
∞	10,827	6,908	5,422	4,617	4,103	3,743	3,475	3,266	3,097	2,959	2,742	2,513	2,266	1,990	1,660	1,447	1,000

Таблица В.12. F-таблица: критические значения для уровня значимости 10%

Знаменатель: степени свободы	Числитель: степени свободы																
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
1	39,863	49,500	53,593	55,833	57,240	58,204	58,906	59,439	59,858	60,195	60,705	61,220	61,740	62,265	62,794	63,061	63,328
2	8,526	9,000	9,162	9,243	9,293	9,326	9,349	9,367	9,381	9,392	9,408	9.425	9,441	9,458	9,475	9,483	9,491
3	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,216	5,200	5,184	5,168	5,151	5,143	5,134
4	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,896	3,870	3,844	3,817	3,790	3,775	3,761
5	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,268	3,238	3,207	3,174	3,140	3,123	3,105
6	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,905	2,871	2,836	2,800	2,762	2,742	2,722
7	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,668	2,632	2,595	2,555	2,514	2,493	2,471
8	3,458	3,113	2,924	2,806	2,726	2,668	2,624	2,589	2,561	2,538	2,502	2,464	2,425	2,383	2,339	2,316	2,293
9	3,360	3,006	2,813	2,693	2,611	2,551	2,505	2,469	2,440	2,416	2,379	2,340	2,298	2,255	2,208	2,184	2,159
10	3,285	2,924	2,728	2,605	2,522	2,461	2,414	2,377	2,347	2,323	2,284	2,244	2,201	2,155	2,107	2,082	2,055
12	3,177	2,897	2,606	2,480	2,394	2,331	2,283	2,245	2,214	2,188	2,147	2,105	2,060	2,011	1,960	1,932	1,904
15	3,073	2,695	2,490	2,361	2,273	2,208	2,158	2,119	2,086	2,059	2,017	1,972	1,924	1,873	1,817	1,787	1,755
20	2,975	2,589	2,380	2,249	2,158	2,091	2,040	1,999	1,965	1,937	1,892	1,845	1,794	1,738	1,677	1,643	1,607
30	2,881	2,489	2,276	2,142	2,049	1,980	1,927	1,884	1,849	1,819	1,773	1,722	1,667	1,606	1,538	1,499	1,456
60	2,791	2,393	2,177	2,041	1,946	1,875	1,819	1,775	1,738	1,707	1,657	1,603	1,543	1,476	1,395	1,348	1,291
120	2,748	2,347	2,130	1,992	1,896	1,824	1,767	1,722	1,684	1,652	1,601	1,545	1,482	1,409	1,320	1,265	1,193
∞	2,706	2,303	2,084	1,945	1,847	1,774	1,717	1,670	1,632	1,599	1,546	1,487	1,421	1,342	1,240	1,169	1,000

Таблица В.13. Ранги для критерия знаков

Размер модифицированной выборки, *n*	Уровень значимости 5%				Уровень значимости 1%			
	Критерий знаков значим, если количество либо				Критерий знаков значим, если количество либо			
	меньше, чем	либо	больше, чем		меньше, чем	либо	больше, чем	
6	1	5	–		–	–	–	
7	1	6	–		–	–	–	
8	1	7	1		1	7	7	
9	2	7	1		1	8	8	
10	2	8	1		1	9	9	
11	2	9	1		1	10	10	
12	3	9	2		2	10	10	
13	3	10	2		2	11	11	
14	3	11	2		2	12	12	
15	4	11	3		3	12	12	
16	4	12	3		3	13	13	
17	5	12	3		3	14	14	
18	5	13	4		4	14	14	
19	5	14	4		4	15	15	
20	6	14	4		4	16	16	
21	6	15	5		5	16	16	
22	6	16	5		5	17	17	
23	7	16	5		5	18	18	
24	7	17	6		6	18	18	
25	8	17	6		6	19	19	
26	8	18	7		7	19	19	
27	8	19	7		7	20	20	
28	9	19	7		7	21	21	
29	9	20	8		8	21	21	
30	10	20	8		8	22	22	
31	10	21	8		8	22	23	
32	10	22	9		9	23	23	

Продолжение табл. В.13

Размер модифицированной выборки, m	Уровень значимости 5%			Уровень значимости 1%		
	Критерий знаков значим, если количество либо			Критерий знаков значим, если количество либо		
	меньше, чем	либо	больше, чем	меньше, чем	либо	больше, чем
33	11		22	9		24
34	11		23	10		24
35	12		23	10		25
36	12		24	10		26
37	13		24	11		26
38	13		25	11		27
39	13		26	12		27
40	14		26	12		28
41	14		27	12		29
42	15		27	13		29
43	15		28	13		30
44	16		28	14		30
45	16		29	14		31
46	16		30	14		32
47	17		30	15		32
48	17		31	15		33
49	18		31	16		33
50	18		32	16		34
51	19		32	16		35
52	19		33	17		35
53	19		34	17		36
54	20		34	18		36
55	20		35	18		37
56	21		35	18		38
57	21		36	19		38
58	22		36	19		39
59	22		37	20		39

Продолжение табл. В.13

Размер модифицированной выборки, m	Уровень значимости 5%			Уровень значимости 1%		
	Критерий знаков значим, если количество либо			Критерий знаков значим, если количество либо		
	меньше, чем	либо	больше, чем	меньше, чем	либо	больше, чем
60	22		38	20		40
61	23		38	21		40
62	23		39	21		41
63	24		39	21		42
64	24		40	22		42
65	25		40	22		43
66	25		41	23		43
67	26		41	23		44
68	26		42	23		45
69	26		43	24		45
70	27		43	24		46
71	27		44	25		46
72	28		44	25		47
73	28		45	26		47
74	29		45	26		48
75	29		46	26		49
76	29		47	27		49
77	30		47	27		50
78	30		48	28		50
79	31		48	28		51
80	31		49	29		51
81	32		49	29		52
82	32		50	29		53
83	33		50	30		53
84	33		51	30		54
85	33		52	31		54

Окончание табл. В.13

Размер модифицированной выборки, m	Уровень значимости 5%			Уровень значимости 1%		
	Критерий знаков значим, если количество либо			Критерий знаков значим, если количество либо		
	меньше, чем	либо	больше, чем	меньше, чем	либо	больше, чем
86	34		52	31		55
87	34		53	32		55
88	35		53	32		56
89	35		54	32		57
90	36		54	33		57
91	36		55	33		58
92	37		55	34		58
93	37		56	34		59
94	38		56	35		59
95	38		57	35		60
96	38		58	35		61
97	39		58	36		61
98	39		59	36		62
99	40		59	37		62
100	40		60	37		63

Таблица В.14. Критические значения для тестов "хи-квадрат"

Степени свободы	Уровень 10%	Уровень 5%	Уровень 1%	Уровень 0,1%
1	2,706	3,841	6,635	10,828
2	4,605	5,991	9,210	13,816
3	6,251	7,815	11,345	16,266
4	7,779	9,488	13,277	18,467
5	9,236	11,071	15,086	29,515
6	10,645	12,592	16,812	22,458
7	12,017	14,067	18,475	24,322
8	13,362	15,507	20,090	26,124
9	14,684	16,919	21,666	27,877
10	15,987	18,307	23,209	29,588

Продолжение табл В.14

Степени свободы	Уровень 10%	Уровень 5%	Уровень 1%	Уровень 0,1%
11	17,275	19,675	24,725	31,264
12	18,549	21,026	26,217	32,909
13	19,812	22,362	27,688	34,528
14	21,064	23,685	29,141	36,123
15	22,307	24,996	30,578	37,697
16	23,542	26,296	32,000	39,252
17	24,769	27,587	33,409	40,790
18	25,989	28,869	34,805	42,312
19	27,204	30,144	36,191	43,820
20	28,412	31,410	37,566	45,315
21	29,615	32,671	38,932	46,797
22	30,813	33,924	40,289	48,268
23	32,007	35,172	41,638	49,728
24	33,196	36,415	42,980	51,179
25	34,382	37,652	44,314	52,620
26	35,563	38,885	45,642	54,052
27	36,741	40,113	46,963	55,476
28	37,916	41,337	48,278	56,892
29	39,087	42,557	49,588	58,301
30	40,256	43,773	50,892	59,703
31	41,422	44,985	52,191	61,098
32	42,585	46,194	53,486	62,487
33	43,745	47,400	54,776	63,870
34	44,903	48,602	56,061	65,247
35	46,059	49,802	57,342	66,619
36	47,212	50,998	58,619	67,985
37	48,363	52,192	59,893	69,346
38	49,513	53,384	61,162	70,703
39	50,660	54,572	62,428	72,055
40	51, 805	55,758	63,691	73,402
41	52,949	56,942	64,950	74,745
42	54,090	58,124	66,206	76,084

Продолжение табл В.14

Степени свободы	Уровень 10%	Уровень 5%	Уровень 1%	Уровень 0,1%
43	55,230	59,304	67,459	77,419
44	56,369	60,481	68,710	78,749
45	57,505	61,656	69,957	80,077
46	58,641	62,830	71,201	81,400
47	59,774	64,001	72,443	82,720
48	60,907	65,171	73,683	84,037
49	62,038	66,339	74,919	85,351
50	63,167	67,505	76,154	86,661
51	64,295	68,669	77,386	87,968
52	65,422	69,832	78,616	89,272
53	66,548	70,993	79,843	90,573
54	67,673	72,153	81,069	91,872
55	68,796	73,311	82,292	93,167
56	69,919	74,468	83,513	94,461
57	71,040	75,624	84,733	95,751
58	72,160	76,778	85,950	97,039
59	73,279	77,931	87,166	98,324
60	74,397	79,082	88,379	99,607
61	75,514	80,232	89,591	100,888
62	76,730	81,381	90,802	102,166
63	77,745	82,589	92,010	103,442
64	78,860	83,675	93,217	104,716
65	79,973	84,821	94,422	105,988
66	81,085	85,965	95,626	107,258
67	82,197	87,108	96,828	108,526
68	83,308	88,250	98,028	109,791
69	84,418	89,391	99,228	111,055
70	85,527	90,531	100,425	112,317
71	86, 635	91,670	101,621	113,577
72	87,743	92,808	102,816	114,835
73	88,850	93,945	104,010	116,091
74	89,956	95,081	105,202	117,346

Окончаниет табл В.14

Степени свободы	Уровень 10%	Уровень 5%	Уровень 1%	Уровень 0,1%
75	91,061	96.217	106,393	118,599
76	92,166	97,351	107,583	119,850
77	93,270	98,484	108,771	121,100
78	94,374	99,617	109,958	122,348
79	95,476	100,749	111,144	123,594
80	96,578	101,879	112,329	124,839
81	97,680	103,010	113,512	126,083
82	98,780	104,139	114,695	127,324
83	99,880	105,267	115,876	127,565
84	100,980	106,395	117,057	129,804
85	102,079	107,522	118,236	131,041
86	103,177	108,648	119,414	132,277
87	104,275	109,773	120,591	133,512
88	105,372	110,898	121,767	134,745
89	106,469	112,022	122,942	135,978
90	107,565	113,145	124,116	137,208
91	108,661	114,268	125,289	138,438
92	109,756	115,390	126,462	139,666
93	110,850	116,511	127,633	140,893
94	111,944	117,632	128,803	142,119
95	113,038	118,752	129,973	143,344
96	114,131	119,871	131,141	144,567
97	115,223	120,990	132,309	145,789
98	116,313	122,108	133,476	147,010
99	117,407	123,225	134,642	148,230
100	118,498	124,342	135,807	149,449

Таблица В.15. Множители для построения $\bar{X}$ - и R -карт

Размер выборки, n	Карта для средних ($\bar{X}$ - карта): Факторы для контрольных границ		Карты для диапазонов (R -карты)				
			Фактор для центральной линии, d_2	Факторы для контрольных границ			
	A	A_2		D_1	D_2	D_3	D_4
2	2,121	1,880	1,128	0	3,686	0	3,267
3	1,732	1,023	1,693	0	4,358	0	2,574
4	1,500	0,729	2,059	0	4,698	0	2,282
5	1,342	0,577	2,326	0	4,918	0	2,114
6	1,225	0,483	2,534	0	5,078	0	2,004
7	1,134	0,419	2,704	0,204	5,204	0,076	1,924
8	1.061	0,373	2,847	0,388	5,306	0,136	1,864
9	1,000	0,337	2,970	0,547	5,393	0,184	1,816
10	0,949	0,308	3,078	0,687	5,469	0,223	1,777
11	0,905	0,285	3,173	0,811	5,535	0,256	1,744
12	0,866	0,266	3,258	0,922	5,594	0,283	1,717
13	0,832	0,249	3,336	1,025	5,647	0,307	1,693
14	0,802	0,235	3,407	1,118	5,696	0,328	1,672
15	0,775	0,223	3,472	1,203	5,741	0,347	1,653
16	0,750	0,212	3,532	1,282	5,782	0,363	1,637
17	0,728	0,203	3,588	1,356	5,820	0,378	1,622
18	0,707	0,194	3,640	1,424	5,856	0,391	1,608
19	0,688	0,187	3,689	1,487	5,891	0,403	1,597
20	0,671	0,180	3,735	1,549	5,921	0,415	1,585
21	0,655	0,173	3,778	1,605	5,951	0,425	1,575
22	0,640	0,167	3,819	1,659	5,979	0,434	1,566
23	0,626	0,162	3,858	1,710	6,006	0,443	1,557
24	0,612	0,157	3,895	1,759	6,031	0,451	1,548
25	0,600	0,153	3,931	1,806	6,056	0,459	1,541

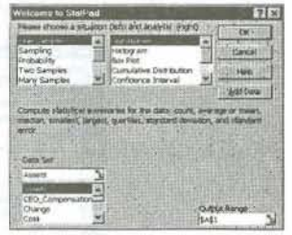
Эти значения взяты из ASTM-STP 15D (American Society for Testing and Materials).

Краткое руководство по применению StatPad

Если у вас установлена надстройка StatPad, то команду для ее запуска вы найдете в меню Excel Tools (Сервис) (или просто откройте файл STATPAD.XLA в Excel). Чтобы начать статистический анализ, просто выберите команду StatPad.

Методы статистического анализа объединены в StatPad в следующие группы.

One Sample	Одна выборка
Two Sample	Две выборки
Multivariate	Многомерный
Sampling	Взятие выборки
Many Sample	Много выборок
Time series	Временные ряды
Probability	Вероятность
Bivariate	Двухмерный
Quality Control	Контроль качества

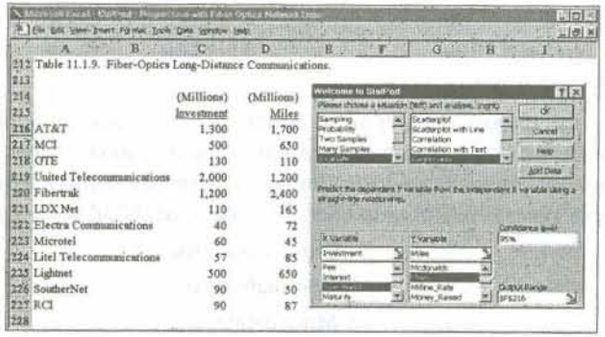


Эти группы представлены в списке слева в главном диалоговом окне StatPad. После того как вы выбрали группу, в списке справа в том же диалоговом окне вы увидите соответствующие способы анализа. После выбора группы и метода анализа в диалоговом окне также появится пояснение, и окно изменится, что позволит вам задать необходимую для выполнения этого анализа информацию (например, доверительный уровень). Вот как выглядит главное диалоговое окно StatPad, когда вы впервые открываете меню Tools в Excel, чтобы вычислить основные характеристики для одномерного набора данных с именем Assets ("Активы")[1] и поместить результаты в конкретную область таблицы, которая задается в окне Output Range.

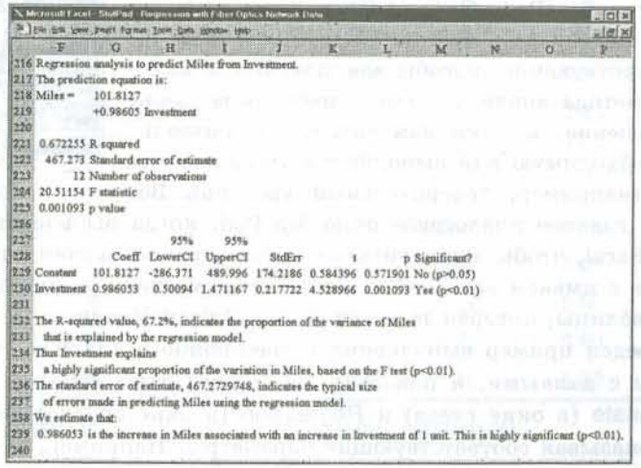
Ниже приведен пример выполнения регрессионного анализа, показывающий начало работы с данными, и показано диалоговое окно StatPad. После выбора ситуации Bivariate (в окне слева) и Regression (в окне справа) диалоговое окно изменится, показывая соответствующие параметры. Например, вы можете опре-

[1] Вы можете либо использовать кнопку *Add Data* в диалоговом окне StatPad, чтобы указать колонку чисел с помощью имени, либо использовать для этой цели имена Excel (например, команду Insert⇨Name⇨Define (Вставка⇨Имя⇨Присвоить)).

делить независимую переменную X (в этом примере "Инвестиции"), зависимую прогнозируемую переменную Y ("Мили"), доверительный уровень (обычно 95, 99 или 99,9%, но можно задать любое число в диапазоне от 0 до 100%). Данные можно выбрать из списков выделенных и названных заранее ("pre-named") наборов данных (или нужно щелкнуть на Add data, чтобы добавить название в списки). Зона вывода (Output range) сообщает StatPad, в какое место вашей рабочей таблицы необходимо поместить вычисленный результат.



После того как вы щелкнете на кнопке Ok в диалоговом окне StatPad, результаты регрессионного анализа появятся в вашей рабочей таблице. Обратите внимание, как StatPad поясняет, что делать с числовыми результатами! R^2 интерпретируется как процент объясненной дисперсии, F-тест сопровождается утверждением о значимости, стандартная ошибка оценки интерпретируется как типичный размер ошибок прогноза, а коэффициент регрессии (с тестом значимости) трактуется как размер увеличения Y, связанного с увеличением X на 1 единицу. Ниже приведены результаты в том виде, как их выводит StatPad.



Ниже приведен список полезных команд StatPad, сгруппированных по главам.

Глава	Группа (левая сторона диалоговой панели *StatPad*)	Метод анализа (правая сторона диалоговой панели *StatPad*)
Глава 3	Одна выборка (One sample)	Гистограмма
Глава 4	Одна выборка (One sample)	Показатели (среднее, медиана, 5 характеристик выборки)
Глава 4	Одна выборка (One sample)	Перцентиль
Глава 4	Одна выборка (One sample)	Блочная диаграмма
Глава 4	Одна выборка (One sample)	Кумулятивное распределение
Глава 4	Одна выборка (One sample)	Перцентильное ранжирование
Глава 5	Одна выборка (One sample)	Показатели (стандартное отклонение)
Глава 7	Вероятность (Probability)	Нормальное распределение вероятности
Глава 7	Вероятность (Probability)	Биномиальное распределение вероятности
Глава 7	Вероятность (Probability)	Биномиальный процент
Глава 8	Взятие выборки (Sampling)	Выборка без замещения
Глава 8	Взятие выборки (Sampling)	Выборка с замещением
Глава 8	Взятие выборки (Sampling)	Различные распределения (нормальное, биномиальное, равномерное)
Глава 8	Одна выборка (One sample)	Показатели (стандартная ошибка)
Глава 9	Одна выборка (One sample)	Доверительный интервал
Глава 9	Две выборки (Two sample)	Доверительный интервал
Глава 10	Одна выборка (One sample)	Проверка гипотез
Глава 10	Две выборки (Two sample)	Проверка гипотез
Глава 11	Двухмерный (Bivariate)	Диаграмма рассеяния, диаграмма рассеяния с линией регрессии
Глава 11	Двухмерный (Bivariate)	Корреляция, значимость корреляции
Глава 11	Двухмерный (Bivariate)	Регрессия
Глава 11	Двухмерный (Bivariate)	Прогноз и остатки
Глава 12	Многомерный (Multivariate)	Диаграммы рассеяния
Глава 12	Многомерный (Multivariate)	Корреляции
Глава 12	Многомерный (Multivariate)	Регрессия
Глава 12	Многомерный (Multivariate)	Прогноз и остатки
Глава 12	Многомерный (Multivariate)	Диаграмма диагностики
Глава 14	Временные ряды (Time series)	Анализ тренда с учетом сезонных колебаний
Глава 15	Много выборок (Many samples)	Характеристики, гистограммы, блочные диаграммы
Глава 15	Много выборок (Many samples)	F-тест (однофакторный дисперсионный анализ ANOVA)
Глава 15	Много выборок (Many samples)	Различие средних (тест наименьшего значимого различия)
Глава 18	Контроль качества (Quality control)	X -столбиковые диаграммы и R -карты
Глава 18	Контроль качества (Quality control)	Процентная карта и карты частот

Ниже приведен перечень ситуаций, способов анализа и пояснений, доступных в StatPad.

Одна выборка

Показатели	Вычисление для данных таких статистических показателей, как частота, среднее, медиана, наименьшее и наибольшее значение, квартили, стандартное отклонение и стандартная ошибка.
Гистограмма	Построение гистограммы для изучения данных, которая показывает форму кривой распределения, типичные значения, изменчивость и выбросы. Данные концентрируются там, где столбики гистограммы самые высокие. Проверьте "настройку параметров программы" для определения оптимальной ширины интервала и точки ориентира.
Блочная диаграмма	Построение блочной диаграммы для изучения данных, которая демонстрирует 5 базовых характеристик (наименьшее и наибольшее значения, нижний и верхний квартили, медиана), а также выбросы.
Кумулятивное распределение	Построение функции кумулятивного распределения данных, которая показывает процент тех значений, которые меньше любого заданного числа. Это дает значения перцентилей.
Доверительный интервал	Вычисление доверительного интервала для среднего генеральной совокупности. Это статистический вывод о генеральной совокупности, основанный на анализе случайной выборки. Можно строить одно- или двухсторонний доверительные интервалы для любого доверительного уровня.
Проверка гипотез	Проверка нулевой гипотезы о том, что среднее генеральной совокупности равно заданному опорному значению. Это статистический вывод о генеральной совокупности, основанный на анализе случайной выборки. Используется двух- или односторонний тест (t-тест Стьюдента).
Перцентиль	Для заданного процента вычисляется значение перцентиля. Для этого значения (перцентиля) в наборе данных существует приблизительно такой процент значений, которые меньше его.
Перцентильное ранжирование	Вычисление перцентильного ранга для заданного значения. Это приблизительный процент таких значений в наборе данных, которые меньше этого заданного значения (перцентиля).

Взятие выборки

Выборка без замещения	Извлечение из большей совокупности случайной выборки без замещения, т.е. ни один элемент не может быть выбран больше одного раза. Все элементы совокупности имеют одинаковые возможности попасть в выборку и все извлекаются независимо друг от друга.
Выборка с замещением	Извлечение из большей совокупности случайной выборки с замещением, т.е. элемент совокупности может быть отобран больше одного раза. Все элементы совокупности имеют одинаковую возможность попасть в выборку и все извлекаются независимо друг от друга.
Равномерное распределение	Отбор случайной выборки из равномерно распределенной совокупности, в которой все значения равновероятно распределены между наименьшим и наибольшим возможными значениями. Полученному результату можно дать имя и использовать его позже.
Нормальное распределение	Отбор случайной выборки из совокупности нормально распределенной с заданными средним и стандартным отклонениями. Полученному результату можно дать имя и использовать его позже.
Биномиальное распределение	Отбор случайной выборки из биномиально распределенной совокупности (количество наступлений события) с заданными количеством испытаний и вероятностью наступления события. Полученному результату можно дать имя и использовать его позже.
Биномиально распределенные проценты (доли)	Отбор случайной выборки из биномиально распределенной совокупности процентов с заданными количеством испытаний и вероятностью наступления события. Полученному результату можно дать имя и использовать его позже.

Вероятность

Нормально распределенная вероятность	Вероятности для нормального распределения: симметричная кривая в форме колокола, заданные среднее и положительное стандартное отклонения.
Биномиально распределенная вероятность	Вероятности для биномиального закона распределения: количество появлений события в результате заданного количества независимых испытаний с заданной вероятностью.
Биномиальный процент	Вероятности для биномиального процента при заданных количестве независимых испытаний и вероятности появления события в каждом испытании.

Две выборки

Показатели	Вычисление статистических обобщающих показателей одномерного распределения для каждого из наборов данных. Также вычисляются разность средних и ее стандартная ошибка. Если размеры выборок одинаковы, то можно указать, что речь идет о двух измерениях для каждого элемента.
Гистограммы	С целью предварительного анализа для каждого набора данных строится гистограмма.

Блочные диаграммы	С использованием одной шкалы измерения (для удобства сравнения) строятся блочные диаграммы для каждого набора данных.
Доверительный интервал	Вычисление доверительного интервала для разности средних генеральных совокупностей. Это статистический вывод. Строится двухсторонний интервал с заданным доверительным уровнем. Если размеры выборок одинаковы, то можно указать, что речь идет о двух измерениях для каждого элемента.
Проверка гипотезы	Проверяется нулевая гипотеза о том, что разница средних генеральных совокупностей равна нулю. Это статистический вывод. Используется двухсторонний t-тест Стьюдента. Если размеры выборок одинаковы, то можно указать, что речь идет о двух измерениях для каждого элемента.

Много выборок

Показатели	Можно выбрать необходимое количество наборов данных. Для каждого набора вычисляются статистические показатели одномерного распределения.
Гистограммы	С целью предварительного анализа данных строятся гистограммы для каждой выборки.
Блочные диаграммы	С использованием одной шкалы измерения (для удобства сравнения) строятся блочные диаграммы для каждого набора данных.
F-тест	Однофакторный дисперсионный анализ (ANOVA). Проверка нулевой гипотезы о том, что средние всех генеральных совокупностей равны. Это статистический вывод.
Различие средних	Доверительный интервал и проверки гипотез о различии средних для каждой пары генеральных совокупностей (тест наименьшего значимого различия). Это статистический вывод.

Двухмерный анализ

Диаграмма рассеяния	Строится диаграмма рассеяния для изучения связи между двумя переменными.
Диаграмма разброса с линией	Для изучения связи между двумя переменными строится диаграмма рассеяния с линией наименьших квадратов.
Корреляция	Сила связи между двумя переменными определяется как обычное число, где 1 указывает на идеальную прямолинейную зависимость, −1 — на идеальную обратно пропорциональную зависимость, а 0 — на отсутствие связи.
Корреляция с проверкой значимости	Определяется и проверяется значимость связи между двумя переменными. Это статистический вывод.

Регрессия	Предсказание с использованием прямолинейной связи значения зависимой переменной Y исходя из независимой переменной X.
Предсказанные значения и разности	Предсказанные на основе X значения Y, остаточная разность (фактическое Y −предсказанное Y) и стандартизованные остатки.
Показатели одномерного распределения	Вычисление показателей одномерного распределения для каждой переменной.
Гистограммы	С целью предварительного анализа данных строится гистограмма для каждой переменной.
Блочные диаграммы	С целью предварительного анализа данных строится блочная диаграмма для каждой переменной.

Многомерный анализ

Диаграмма рассеяния	Выберите любое необходимое количество X-переменных, но только одну Y-переменную. Построение диаграмм рассеяния для каждой пары переменных с целью изучения связей между ними.
Корреляция	Вычисляется сила связи между парами переменных в виде матрицы коэффициентов корреляции (1 означает идеальную положительную корреляцию, −1 — идеальную отрицательную корреляцию, а 0 — отсутствие связи).
Регрессия	Предсказание значения зависимой переменной Y по независимым X-переменным исходя из линейной связи.
Предсказанные значения и разности	Предсказанные исходя из X-переменных значения Y, остаточные разности (фактическое Y −предсказанное Y) и стандартизованные остатки.
Диаграмма диагностики	Поиск проблем в регрессионной линейной модели, таких как неравная изменчивость или нелинейность.
Показатели одномерного анализа	Вычисление одномерных обобщающих показателей для каждой переменной.
Гистограммы	С целью предварительного анализа данных строятся гистограммы для каждой переменной.
Блочные диаграммы	С целью предварительного анализа данных строятся блочные диаграммы для каждой переменной.

Временны*е* ряды

Тренд с сезонными колебаниями	Разложение на (1) долгосрочный тренд (линейный или экспоненциальный), (2) повторяющаяся сезонная компонента (месячная или квартальная), (3) блуждающая циклическая компонента, (4) несистематическая случайная компонента. Включает поправку на сезонные колебания и прогноз.

Контроль качества

X-столбиковые диаграммы и R-карты	На карте изображаются средние значения и размах данных, что позволяет решить, находится ли процесс под контролем. Размер подгруппы можно выбирать от 2 до 25. Можно установить значение нормы, если оно известно.
Карта процентов и частот	Карта процентов или частот позволяет увидеть, находится ли процесс под контролем. Данными могут быть либо частоты, либо проценты (частоты делят на размер выборки). Можно установить значение нормы, если оно известно.

Словарь терминов

A

ARIMA-процесс Бокса–Дженкинса (Box-Jenkins ARIMA process). Одна из целого семейства линейных статистических моделей, основанная на нормальном распределении и позволяющая имитировать поведение множества различных временных рядов, комбинируя процессы авторегрессии (AR), интегрированные (I) процессы и процессы скользящего среднего (MA).

F

F-статистика (F statistic). Основа для F-теста в дисперсионном анализе; представляет собой отношение двух измерений дисперсии, используемых при выполнении каждой проверки гипотез.

F-таблица (F table). Содержит критические значения распределения F-статистики при условии справедливости нулевой гипотезы; таким образом, при условии справедливости нулевой гипотезы значение F-статистики превосходит критическое значение лишь в некотором определенном проценте случаев (например, в 5% случаев).

F-тест (множественная регрессия) (F table [multiple regression]). Общий тест, позволяющий выяснить, объясняют ли X-переменные значимую долю вариации переменной Y.

F-тест [ANOVA] (F test [ANOVA]). Основанный на F-статистике тест, который используется при тестировании каждой из гипотез в дисперсионном анализе.

P

***p*-значение** (*p*-value). Свидетельствует о том, насколько неожиданным является факт, что данные соответствуют нулевой гипотезе. Малые *p*-значения обозначают большую неожиданность такого факта и приводят к отказу от нулевой гипотезы H_0. Принято отвергать H_0, когда *p*-значение меньше, чем 0,05.

R

R^2. См. *коэффициент детерминации* (coefficient of determination).

***R*-диаграмма для контроля качества** (*R*-chart, for quality control). Содержит диапазон (наибольшее значение минус наименьшее значение) для каждой выборки — наряду с соответствующей центральной линией и контрольными границами, — что позволяет отслеживать изменчивость процесса.

T

t-статистика (t statistic). Один из способов выполнения t-теста:

$$t = (\bar{X} - \mu_0) / S_{\bar{X}}.$$

t-таблица (t table). Таблица, которая вносит поправку на добавленную неопределенность, вызванную тем, что вместо неизвестного точного значения изменчивости для исследуемой генеральной совокупности используется некоторая ее оценка (стандартная ошибка).

t-тест для зависимых выборок (paired t test). Используется в случаях, когда требуется выяснить, соответствуют ли двум выборкам одинаковые генеральные средние, если между этими двумя выборками можно установить естественное соответствие (например, измерения "до" и "после" некоторого воздействия, выполненные для той же группы людей).

t-тест, или t-тест Стьюдента (t test or Student's t test). Проверка гипотезы о среднем значении.

t-тесты для отдельных коэффициентов регрессии для множественной регрессии (t tests for individual regression coefficients, for multiple regression). Если регрессия значима, метод дальнейшего статистического вывода относительно отдельных коэффициентов регрессии.

U

U-тест Манна–Уитни (Mann-Whitney U test). Непараметрический тест для двух независимых выборок.

X

$\bar{X}$-диаграмма для управления качеством ($\bar{X}$ chart, for quality control). Представление среднего значения каждой выборки вместе с соответствующей центральной линией и контрольными границами, что позволяет отслеживать уровень процесса.

A

Анализ и методы (analysis and methods). Раздел отчета, в котором вы интерпретируете данные путем представления графиков, выводов и результатов с соответствующими комментариями и пояснениями.

Анализ сезонных тенденций (trend-seasonal analysis). Непосредственный, интуитивный подход к оценке четырех базовых компонентов месячного или квартального временного ряда: долгосрочная тенденция, сезонные особенности, циклическая вариация и нерегулярный компонент.

Асимметричное (скошенное) распределение (skewed distribution). Распределение, которое не является ни симметричным, ни нормальным, поскольку кривая распределения приближается к горизонтальной оси достаточно резко с одной стороны и более плавно — с другой.

Б

Байесовский анализ (Bayesian analysis). Статистические методы, использующие формальным математическим способом субъективные вероятности.

Безусловная вероятность (unconditional probability). Обычная (без поправок) вероятность.

Бимодальное распределение (bimodal distribution). На этот тип распределения указывает наличие на гистограмме двух четко выраженных отдельных групп.

Биномиальное доля (binomial proportion). Доля $\rho = X/n$, которая также представляет и процент.

Биномиальное распределение (binomial distribution). Распределение случайной переменной X, которая представляет количество наступлений определенного события в серии из n испытаний при условии, что для каждого из n испытаний указанное событие всегда имеет одну и ту же вероятность наступления (π) и что эти испытания не зависят друг от друга.

Блочная диаграмма (box plot). Диаграмма, представляющая графически пять основных характеристик распределения.

Большое среднее для ANOVA (grand average, for ANOVA). Среднее всех значений данных из объединения всех выборок.

B

Введение (introduction). Несколько абзацев, расположенных ближе к началу отчета, в которых описывают предысторию, исследуемые вопросы и те данные, которые будут использоваться.

Вероятность (probability). Возможность наступления каждого из множества потенциальных будущих событий, которая основывается на совокупности предположений относительно того, как устроен окружающий нас мир.

Вероятность события (probability of an event). Число в диапазоне от 0 до 1, которое характеризует возможность наступления события каждый раз, когда выполняется соответствующий случайный эксперимент.

Взаимодействие для множественной регрессии (interaction, for multiple regression). Такая взаимосвязь между двумя X-переменными и переменной Y, при которой изменение обеих этих переменных вызывает ожидаемое изменение Y, отличное от суммы изменений Y, полученных в результате изменения каждой из X-переменных отдельно.

Взаимоисключающие события (mutually exclusive events). Два события, которые не могут произойти одновременно.

Взвешенное среднее (weighted average). Мера, подобная среднему значению, за исключением того, что она позволяет назначать индивидуальную значимость (важность, или *вес*) каждому из элементов данных.

Внутривыборочная (внутригрупповая) изменчивость для ANOVA (within-sample variability, for ANOVA). Полная мера изменчивости каждой из выборок.

Вторичные данные (secondary data). Данные, собранные ранее кем-то другим для собственных целей.

Выбор переменных (variable selection). Эта задача возникает тогда, когда имеется внушительный перечень потенциально полезных объясняющих X-переменных и необходимо принять решение о том, какие именно из этих переменных следует включить в уравнение регрессии. Когда X-переменных слишком много, то качество результатов снижается, поскольку информация неэффективно расходуется на оценивание бесполезных параметров. С другой стороны, отсутствие одной или нескольких важных X-переменных приводит к снижению качества прогнозов вследствие потери нужной информации.

Выборка (sample). Меньшая совокупность единиц, извлеченная из генеральной совокупности.

Выборка без замещения (sampling without replacement). Схема построения выборки, в которой любая единица генеральной совокупности не может попасть в выборку более одного раза.

Выборка с замещением (sampling with replacement). Схема построения выборки, в которой любая единица генеральной совокупности может попасть в выборку более одного раза.

Выборочная статистика (sample statistic). Какой-либо показатель, вычисленный на основе данных выборки.

Выборочное пространство (sample space). Перечень всех возможных исходов (результатов) случайного эксперимента, составленный заранее, когда еще неизвестно, что произойдет при выполнении эксперимента.

Выборочное распределение (sampling distribution). Распределение вероятностей результатов каких-либо измерений, полученное на основе случайной выборки данных.

Выборочное стандартное отклонение (sample standard deviation). Представляет собой меру изменчивости и используется для обобщающего перехода от имеющихся данных к некоторой более крупной генеральной совокупности (реальной или гипотетической).

Выброс (сильно отличающееся значение) (outlier). Значение, которое, по-видимому, не согласуется с другими значениями данных, являясь либо слишком большим, либо слишком малым.

Выброс (сильно отличающееся значение) в двумерных данных (bivariate outlier). Точка на диаграмме рассеяния, "выпадающая" из общего характера взаимосвязи в данных.

Выводы и заключение (conclusion and summary). Раздел отчета, в котором, в завершение всего сказанного, дается "общая картина", включающая все наиболее важные мысли, о которых вы хотели бы еще раз напомнить читателям.

Вычисление среднего значения (average). Общепринятый метод нахождения типичного значения для некоторой совокупности чисел. Это типичное значение вычисляют путем сложения всех чисел совокупности и деления полученной суммы на число элементов совокупности; иногда этот показатель называют просто *средним* (mean).

Г

Генеральная совокупность (population). Изучаемая совокупность некоторых единиц (людей, объектов или чего-либо другого).

Гипотеза (hypothesis). Утверждение относительно генеральной совокупности, которое может быть верным или неверным; данные помогают принять решение о том, какую одну из двух гипотез можно считать истинной.

Гистограмма (histogram). Изображение частот значений в виде совокупности столбцов, возвышающихся над числовой линией; указывает, как часто в совокупности данных встречаются те или иные значения.

Гистограмма типа "ствол и листья" (stem-and-leaf histogram). Гистограмма, столбцы которой формируются путем записи чисел одних над другими.

Д

Данные временного ряда (time-series data). Значения данных, которые фиксируются в определенной, имеющей содержательный смысл, последовательности.

Данные об одном временном срезе (cross-sectional data). Набор данных, в котором порядок записи данных не имеет никакого значения.

Двумерные данные (bivariate data). Набор данных, в котором для каждого элемента указана информация о двух некоторых свойствах.

Двусторонний тест (two-sided test). Тест, для которого исследовательская гипотеза допускает возможность того, что параметр генеральной совокупности может находиться по обе стороны от эталонного значения.

Дерево вероятностей (probability tree). Рисунок, содержащий вероятности и некоторые условные вероятности для сочетаний из двух и более событий.

Диагностическая диаграмма (множественная регрессия) (diagnostic plot [multiple regression]). Диаграмма рассеяния для значений ошибок предсказания (остатков) как функции от прогнозируемых значений; используется для поиска в данных различных проблем, требующих решения.

Диаграмма Венна (Venn diagram). Рисунок, на котором все множество возможных исходов (выборочное пространство) изображено в виде внешнего прямоугольника, внутри которого расположены события, часто в виде кружков или овалов.

Диаграмма Парето (Pareto diagram). Представление причин различных дефектов в порядке от наиболее часто встречающихся к наименее часто встречающимся, что позволяет сосредоточить внимание на наиболее важных проблемах.

Диаграмма рассеяния (scatterplot). Графическое представление, позволяющее исследовать двумерные данные (Y как функция от X); дает возможность получить наглядную картину взаимосвязи в исследуемых данных.

Дискретная количественная переменная (discrete quantitative variable). Переменная, которая может принимать значения только из определенного перечня возможных чисел (например, 0 или 1, или из списка 0, 1, 2, 3,...).

Дискретная случайная переменная (discrete random variable). Случайная переменная, для которой можно перечислить все ее возможные значения.

Дисперсионный анализ (analysis of variance — ANOVA). Общая модель проверки статистических гипотез, основанная на тщательном анализе различных источников изменчивости в той или иной сложной ситуации.

Дисперсия (variance). Квадрат стандартного отклонения. Этот показатель несет ту же информацию, что и стандартное отклонение, но труднее интерпретируется, поскольку единицы измерения дисперсии представляют собой единицы измерения исходных данных, возведенные в квадрат (например, доллары в квадрате, мили в квадрате на каждый галлон в квадрате, килограммы в квадрате и другие малопонятные вещи).

Доверительный интервал (confidence interval). Вычисленный на основе данных интервал, который с заданной вероятностью включает интересующий нас (неизвестный) параметр генеральной совокупности.

Доверительный уровень (confidence level). Вероятность попадания параметра генеральной совокупности в доверительный интервал; традиционно устанавливается равным 95%, хотя часто используются также уровни 90, 99 и 99,9%.

Дополнение [NOT] (complement [NOT]). Альтернативное событие, которое происходит лишь в том случае, когда изучаемое событие *не* происходит.

З

Зависимые события (dependent events). Такие события, для которых информация об одном из них влияет на оценку вероятности другого.

Закон больших чисел (law of large numbers). Правило, утверждающее, что относительная частота (случайная величина) будет приближаться к вероятности (точному, определенному числу), если эксперимент будет выполняться многократно.

И

Изменчивость, разнообразие, неопределенность, разброс или размах (variability, diversity, uncertainty, dispersion, or spread). Степень отличия значений данных друг от друга.

Индикаторная переменная (indicator variable). Называется также *фиктивной переменной* (dummy variable); количественная переменная, которая может принимать только два значения — 0 и 1 — и используется как объясняющая X-переменная для представления качественных категориальных данных.

Интервал прогнозирования (prediction interval). Дает возможность с заданной вероятностью использовать данные из некоторой выборки для прогнозирования нового наблюдения, что позволяет получить это дополнительное наблюдение таким же образом, как были получены данные.

Исследование данных (exploring the data). Изучение имеющейся совокупности данных с различных точек зрения, описание данных и их обобщение.

Исследовательская гипотеза, или альтернативная гипотеза (research hypothesis or alternative hypothesis). Гипотеза, которую необходимо доказывать, т.е. для принятия которой необходимы убедительные аргументы против H_0 ; обозначается H_1.

Исход (outcome). Результат выполнения случайного эксперимента, описывающий и фиксирующий наблюдаемые последствия.

К

Качественная переменная (qualitative variable). Переменная, которая указывает, в какую из нескольких нечисловых категорий попадает элемент.

Квартили (quartiles). 25-й и 75-й перцентили.

Кластеринг, или группировка (clustering). Может иметь место в двумерных данных в том случае, когда на диаграмме рассеяния присутствуют хорошо отличающиеся друг от друга группы точек; в подобных случаях бывает необходимо анализировать каждую группу по отдельности.

Ковариация X и Y (covariance X and Y). Числитель в формуле вычисления коэффициента корреляции.

Количественная переменная (quantitative variable). Переменная, значения которой представляют собой имеющие содержательный смысл числа.

Контрольная карта (control chart). Отображение последовательных измерений некоторого процесса вместе с центральной линией и контрольными границами, которые помогают понять, вышел процесс из под контроля или нет.

Корреляционная матрица (correlation matrix). Таблица, содержащая коэффициенты корреляции для каждой пары переменных из многомерной совокупности данных.

Коэффициент вариации (coefficient of variation). Стандартное отклонение, деленное на среднее значение; характеризует относительную изменчивость данных, выраженную как процент от среднего значения.

Коэффициент детерминации, R^2 (двумерные данные) (coefficient of determination, R^2 [bivariate data]). Квадрат корреляции, показывающий, какой процент вариации переменной Y объясняется переменной X.

Коэффициент детерминации, R^2 (множественная регрессия) (coefficient of determination, R^2 [multiple regression]). Мера, которая интерпретируется как процент вариации переменной Y, который объясняется или может быть отнесен к X-переменным.

Коэффициент корреляции, r (correlation coefficient, r). Число в диапазоне от -1 до 1, характеризующее силу линейной взаимосвязи.

Коэффициент поправки на конечность генеральной совокупности (finite-population correction factor). Вводится в формулу вычисления стандартной ошибки, когда соответствующая генеральная совокупность мала и выборка является существенной частью генеральной совокупности.

(коэффициент поправки на конечность генеральной совокупности) $\times$
$\times$ (стандартная ошибка) =

$$= \sqrt{\frac{N-n}{N}} S_{\bar{x}} = \sqrt{\frac{N-n}{N}} \frac{S}{\sqrt{n}}.$$

Коэффициент регрессии для множественной регрессии (regression coefficient, for multiple regression). Коэффициент b_j для j-й X-переменной, показывающий влияние X_j на Y после внесения поправки на другие X-переменные; коэффициент b_j указывает, на какое увеличение Y можно рассчитывать в случае, которое

ничем не отличается от нынешнего за исключением того, что значение X_j увеличилось на единицу.

Коэффициент регрессии, b, Y на X, для двумерных данных (regression coefficient, b, of Y on X, for bivariate data). Наклон линии наименьших квадратов.

Крайние (экстремальные) значения (extremes). Наименьшее и наибольшее значения, зачастую представляющие особый интерес.

Критерий знаков (sign test). Исходя из количества таких значений в выборке, которые меньше эталонного значения, позволяет выяснить, равняется ли значение медианы генеральной совокупности заданному эталонному значению.

Критерий знаков для разностей (sign test for the differences). Тест для разностей (или изменений), когда приходится иметь дело со связанными наблюдениями (например, измерения "до" и "после" некоторого воздействия); непараметрическая процедура, позволяющая проверить, существенно ли различаются между собой два столбца данных.

Критическое t-значение (critical t value). t-значение из t-таблицы, которое используется в t-тесте.

Критическое значение (critical value). Соответствующее значение, которое берут из подходящей стандартной статистической таблицы для сравнения со значением тестовой статистики.

Л

Линейная взаимосвязь в двумерных данных (linear relationship, in bivariate data). Наблюдается тогда, когда точки на диаграмме рассеяния расположены произвольным образом с постоянным разбросом вокруг прямой линии.

Линейная модель (linear model). Модель, исходящая из того, что наблюдаемое значение Y определяется линейными соотношениями в генеральной совокупности плюс нормально распределенная случайная ошибка.

Линейный регрессионный анализ для двумерных данных (linear regression analysis, for bivariate data). Прогнозирование с помощью прямой линии значения одной переменной исходя из значений другой переменной.

Линия наименьших квадратов, $Y = a + bX$ (least-square line, $Y = a + bX$). Линия, характеризующаяся наименьшей (из всех возможных линий) суммой квадратов ошибок прогноза по вертикальной оси; используется в качестве наилучшей линии прогноза, основанного на имеющихся данных.

Логарифм (logarithm). Часто используется для преобразования асимметрии в симметрию, поскольку позволяет растянуть шкалу вблизи нуля и распределить большое количество близко расположенных малых значений.

Ложная корреляция (spurious correlation). Высокая корреляция, которая на самом деле объясняется действием некоторого третьего фактора.

М

Медиана (median). Значение, расположенное посередине ряда таким образом, что делит данные на две половины, одна из которых меньше этого значения, а другая — больше.

Межвыборочная (межгрупповая) вариация (between-sample variability). Мера отличия друг от друга средних значений выборок; используется в дисперсионном анализе (ANOVA).

Многомерные данные (multivariate data). Совокупности данных, которые содержат результат измерения трех (или более) свойств для каждого элемента.

Множественная регрессия (multiple regression). Прогнозирование значений одной переменной Y по двум или нескольким X-переменным.

Мода (mode). Наиболее распространенная категория; значение, чаще всего встречающееся в совокупности данных.

Модель множественной линейной регрессии (multiple regression linear model). Модель, в которой предполагается, что наблюдаемое значение Y определяется линейными соотношениями в генеральной совокупности плюс независимые нормально распределенные случайные ошибки.

Модель, математическая модель, или процесс [временные ряды] (model, mathematical model, or process [time series]). Система уравнений, которая позволяет получать искусственные наборы данных типа временных рядов.

Модифицированный размер выборки для критерия знаков (modified sample size, for the sign test). Количество m значений данных, которые отличаются от эталонного значения θ_0.

Мультиколлинеарность (multicollinearity). Проблема, возникающая в случае, когда некоторые из объясняющих (X) переменных слишком подобны между собой. В таком случае трудно получить качественные оценки отдельных коэффициентов регрессии по причине нехватки информации для принятия решения относительно того, какую (или какие) из этих переменных необходимо использовать для объяснения.

Н

Наблюдение случайной переменной (observation of random variable). Фактическое значение случайной переменной.

Набор данных (data set). Совокупность, содержащая результаты измерения (или измерений) для каждого из элементов, причем для всех элементов измеряются либо одно и то же свойство, либо одни и те же наборы свойств.

Наклон, b (slope, b). Количество единиц измерения Y, приходящихся на одну единицу X; указывает крутизну подъема (или снижения, если b — отрицательное число) соответствующей прямой линии.

Независимость двух качественных переменных (independence of two qualitative variables). Отсутствие взаимосвязи между двумя качественными переменными, когда знание значения одной переменной не помогает прогнозировать значение другой переменной; другими словами, вероятности категорий одной пере-

менной равны соответствующим условным вероятности при *заданном* значении другой переменной.

Независимые события (independent events). Два события, для которых информация об одном событии не влияет на оценку вероятности наступления другого события.

Независимый *t*-тест (unpaired *t* test). Используется для проверки того, характеризуются ли две заданные выборки одинаковыми средними значениями генеральной совокупности в случае отсутствия естественной связи между парами элементов этих двух выборок (т.е. каждая из них является независимой выборкой из своей собственной генеральной совокупности).

Нелинейная взаимосвязь в двумерных данных (nonlinear relationship, in bivariate data). Характеризуется графиком, на котором точки группируются не вокруг прямой, а вокруг некоторой кривой линии.

Непараметрические методы (nonparametric methods). Статистические процедуры для проверки гипотез, которые не требуют нормального распределения (или какой-либо другой конкретной формы распределения), поскольку основываются на частотах или рангах, а не на фактических значениях данных.

Неправильный выбор модели (model misspecification). Означает наличие множества потенциальных несоответствий между вашим приложением и моделью множественной линейной регрессии. Исследование данных часто позволяет выявить ряд потенциальных проблем, связанных с нелинейностью, неравной изменчивостью или наличием резко выделяющихся значений. Это может приводить или не приводить к возникновению реальных проблем: даже если гистограммы некоторых переменных асимметричны и даже если некоторые диаграммы рассеяния не линейны, модель множественной линейной регрессии может быть, тем не менее, вполне применимой. Диагностическая диаграмма может помочь понять, действительно ли проблема настолько серьезна, что требует решения.

Непрерывная количественная переменная (continuous quantitative variable). Любая количественная переменная, которая не является дискретной, т.е. не может быть ограничена простым перечнем возможных значений.

Непрерывная случайная переменная (continuous random variable). Случайная переменная, которая может принимать любые значения из некоторого диапазона.

Неравная изменчивость в двумерных данных (unequal variability, in bivariate data). Проблема в данных, характеризующаяся тем, что при перемещении по диаграмме рассеяния в горизонтальном направлении резко возрастает изменчивость по вертикали; это приводит к тому, что корреляционный анализ и регрессионный анализ становятся ненадежными. Эти проблемы можно устранить, либо воспользовавшись соответствующими преобразованиями, либо используя так называемую взвешенную регрессию.

Нерегулярный компонент (irregular component). Краткосрочный, случайный компонент временного ряда, представляющий остаточную вариацию, которая не может быть объяснена.

Неслучайная причина изменения (assignable cause of variation). Основание того, *почему* возникла соответствующая проблема, в тех случаях, когда этому можно найти разумное объяснение.

Несмещенная оценка (unbiased estimator). Оценка, которая корректна в среднем, т.е. не является систематически завышенной или заниженной в сравнении с соответствующим параметром генеральной совокупности.

Нестационарный процесс (nonstationary process). Процесс, который со временем все больше и больше удаляется от своего исходного состояния.

Номинальные данные (nominal data). Категории качественной переменной, которым не присуща естественная, содержательно обоснованная упорядоченность.

Нормальное распределение [данные] (normal distribution [data]). Определенного типа, идеализированная, гладкая колоколообразная гистограмма без какой-либо случайности.

Нормальное распределение [случайная переменная] (normal distribution [random variable]). Непрерывное распределение, представленное хорошо известной колоколообразной кривой.

Нулевая гипотеза (null hypothesis). Гипотеза, которую принимают по умолчанию и обозначают H_0; зачастую указывает какой-либо особый случай (например, чистую случайность).

О

Обобщение (summarization). Использование одного или нескольких отобранных или вычисленных значений для представления набора данных.

Общее резюме (executive summary). Абзац в самом начале отчета; в этом абзаце описываются наиболее важные факты и выводы из вашей работы.

Объединение [*или*] (union [*or*]). Событие, которое происходит каждый раз, когда в результате однократного выполнения случайного эксперимента происходит одно событие *или* другое событие (или оба эти события вместе).

Оглавление (table of contents). Раздел отчета, который следует за общим резюме и содержит перечень названий разделов отчета с указанием соответствующих страниц.

Одномерные данные (univariate data). Наборы данных, в которых для каждого элемента указывается только одна порция информации.

Односторонний *t*-тест (one-sided *t* test). *t*-тест, который используется в случае, когда нулевая гипотеза утверждает, что μ находится по одну сторону от μ_0, а альтернативная гипотеза утверждает, что μ находится по другую сторону от μ_0.

Односторонний доверительный интервал (one-sided confidence interval). Определение — с известной степенью доверия — интервала, такого, что среднее в генеральной совокупности либо не меньше, либо не больше, чем некоторое вычисленное значение.

Однофакторный дисперсионный анализ (one-way analysis of variance). Используется для проверки значимости различий нескольких средних, полученных в разных ситуациях.

Ожидаемое значение или среднее случайной переменной (expected value or mean of a random variable). Типичное или среднее значение случайной переменной.

Основа выборки (frame). Схема, позволяющая получить доступ к элементам генеральной совокупности по номерам от 1 до N (размер генеральной совокупности).

Остаток для двумерных данных (residual, for bivariate data). Ошибка прогнозирования для каждой из точек данных; указывает на то, насколько далеко от линии (т.е. выше или ниже ее) находится соответствующая точка.

Отклонение (deviation). Расстояние между отдельным значением и средним.

Относительная частота (relative frequency). При многократном повторении случайного эксперимента — доля (случайная величина), которую составляет количество появлений определенного события в общем количестве повторений случайного эксперимента.

Отношение к скользящему среднему (ratio-to-moving-average). Метод, в соответствии с которым производится деление ряда на гладкое скользящее среднее, что необходимо для анализа сезонных трендов временных рядов.

Отсутствие взаимосвязи в двумерных данных (no relationship, in bivariate data). Соответствует чисто случайному расположению точек на диаграмме рассеяния, которое характеризуется отсутствием выраженной тенденции к наклону вверх или вниз (при перемещении по диаграмме слева направо).

Оценивание неизвестного значения (estimating an unknown quantity). Наиболее обоснованное предположение о значении, которое можно сделать исходя из имеющихся данных.

Оценка (estimate). Некоторое число, вычисленное на основе данных.

Ошибка второго рода (type II error). Происходит в случае, когда верна альтернативная (исследовательская) гипотеза, но вместо нее принимают нулевую гипотезу и объявляют, что результат *не* является статистически значимым.

Ошибка оценки (error of estimation). Разность между значением параметра генеральной совокупности и его статистической оценкой (значением статистики, используемой для оценивания этого параметра); как правило, неизвестна.

Ошибка первого рода (type I error). Происходит в случае, когда нулевая гипотеза верна, но она отвергается и результат объявляется статистически значимым.

Ошибки прогнозирования или остатки для множественной регрессии (prediction errors or residuals, for multiple regression). Вычисляются как Y – (прогнозируемое Y).

П

Параметр (parameter). Любой показатель, вычисленный для всей генеральной совокупности.

Параметр генеральной совокупности (population parameter). Какое-либо число, вычисленное для всей генеральной совокупности.

Параметрические методы (parametric methods). Статистические процедуры, для которых необходима полностью определенная модель.

Первичные данные (primary data). Данные, полученные в том случае, когда вы разработали план сбора данных (даже если работа по сбору данных выполнялась другими).

Переменная (variable). Порция информации, указанная для каждого элемента (например, его стоимость).

Перепись (census). Выборка, которая включает всю генеральную совокупность, так что $n = N$.

Пересечение [*и*] (intersection [*and*]). Событие, которое происходит тогда, когда в результате однократного выполнения случайного эксперимента происходят одно событие *и* другое событие.

Перцентиль (percentile). Обобщающие характеристики, представляющие ранги в виде процентов (в диапазоне от 0 до 100), а не в виде чисел от 1 до n; при этом нулевой перцентиль соответствует наименьшему числу, 100-й перцентиль — наибольшему числу, 50-й перцентиль — медиане и т.д.

Пилотное исследование (pilot study). Сокращенный вариант исследования, призванный помочь вам выявить те или иные проблемы (если таковые существуют) и устранить их, прежде чем вы приступите к проведению полномасштабного исследования.

Планирование исследования (designing the study). Фаза, включающая планирование конкретных деталей сбора данных, например использование случайной выборки из некоторой генеральной совокупности.

Подробная блочная диаграмма (detailed box plot). Блочная диаграмма со специально помеченными выбросами данных, а также с достаточно сильно отклоняющимися наблюдениями, которые не отнесены к выбросам.

Полиномиальная регрессия (polynomial regression). Один из способов решения проблемы нелинейности, при котором Y прогнозируется на основе одной переменной X и ряда степеней этой переменной (X^2, X^3 и т.д.).

Порядковые данные (ordinal data). Категории качественной переменной, которым присуща естественная, содержательно обоснованная упорядоченность.

Постоянный член, *а* (двумерные данные) (constant term, *a* [bivariate data]). Отрезок, отсекаемый линией наименьших квадратов на вертикальной оси.

Пределы прогноза (forecast limits). Доверительные пределы прогноза (если соответствующая модель позволяет вычислить такие пределы); если эта модель соответствует вашим данным, то будущее наблюдение с вероятностью, например, 95% окажется в этих пределах.

Предположения для построения доверительного интервала (assumptions for confidence interval). (1) Соответствующие данные представляют собой случайную выборку из рассматриваемой генеральной совокупности; (2) измеряемая величина имеет (приблизительно) нормальное распределение.

Предположения для проверки статистических гипотез (assumptions for hypothesis testing). (1) Соответствующая совокупность данных представляет собой случайную выборку из рассматриваемой генеральной совокупности; (2) измеряемая величина имеет (приблизительно) нормальное распределение.

Преобразование (transformation). Замена каждого значения данных другим числом (например, логарифмом этого значения), облегчающим проведение статистического анализа.

Приложение (appendix). Раздел отчета, который содержит весь вспомогательный материал, достаточно важный, чтобы приложить его к отчету, но не достаточно важный для того, чтобы включить его в основной текст отчета.

Проверка гипотез (hypothesis testing). Использование данных для выбора одной из двух (или нескольких) возможностей с целью решить вопрос относительно некоторой неопределенной ситуации; эта процедура часто используется для того, чтобы отличить структуру от простой случайности, и должна рассматриваться как полезный инструмент при принятии исполнительных решений.

Проверочная статистика (test statistic). Наиболее полезный показатель, который можно вычислить на основе имеющихся данных с целью принятия решения о справедливости одной из двух альтернативных указанных гипотез.

Прогноз для временных рядов (forecast, for time series). Ожидаемое (т.е. среднее) значение характеристики будущего поведения оцениваемой модели.

Прогнозируемое значение для двумерных данных (predicted value, for bivariate data). Прогноз Y при заданном значении X; определяется путем подстановки значения X в уравнение линии наименьших квадратов.

Процентная диаграмма для контроля качества (percentage chart, for quality control). Графическое изображение процента дефектов — наряду с соответствующей центральной линией и контрольными границами, — позволяющее отслеживать частоту появления дефектных элементов в результате работы изучаемого процесса.

Процесс авторегрессии (autoregressive [AR] process). Процесс, в котором каждое наблюдение состоит из линейной функции предшествующего наблюдения плюс независимый случайный шум.

Процесс авторегрессии и интегрированного скользящего среднего (autoregressive integrated moving-average [ARIMA] process). Процесс, в котором изменения или различия являются результатом процесса авторегрессии и скользящего среднего (autoregressive moving-average [ARMA] process).

Процесс авторегрессии и скользящего среднего (autoregressive moving-average [ARMA] process). Процесс, в котором каждое наблюдение состоит из линейной функции предшествующего наблюдения плюс независимый случайный шум минус определенная доля предшествующего случайного шума.

Процесс для контроля качества (process, for quality control). Любая экономическая деятельность, которая преобразует на выходе то, что поступает на ее вход, в определенные результаты.

Процесс скользящего среднего (moving-average [MA] process). Процесс, каждое наблюдение в котором состоит из константы, μ (долгосрочное среднее значение процесса), плюс независимый случайный шум минус определенная доля предшествующего случайного шума.

Процесс случайного шума (random noise process). Случайная выборка (независимые наблюдения) из нормально распределенной совокупности с постоянными значениями среднего и стандартного отклонения.

Пять основных показателей распределения (five-number summary). Основные обобщающие характеристики совокупности данных: наименьшее значение, нижний квартиль, медиана, верхний квартиль и наибольшее значение.

P

Размах (диапазон) (range). Результат вычитания наименьшего значения набора данных из наибольшего; характеризует размер или протяженность набора данных.

Ранг (rank). Широко используется в непараметрических статистических методах и представляет собой позицию значения в наборе данных после того, как набор упорядочен. Каждому из чисел 1, 2, 3,... , n ставится в соответствие определенное значение из набора данных таким образом, что наименьшее значение имеет ранг 1, следующее наименьшее (но больше его) значение имеет ранг 2, и так до самого большого значения данных, имеющего ранг n.

Распределение вероятностей (probability distribution). Модель вероятностей для некоторой случайной переменной.

Распределение Пуассона (Poisson distribution). Распределение дискретной случайной переменной, для которой соответствующие события появляются независимо и случайно во времени, а средняя доля наступления события постоянна во времени.

Регрессионный анализ (regression analysis). Прогнозирование одной Y-переменной по одной или нескольким X-переменным.

Репрезентативная выборка (representative sample). Выборка, в которой каждая характеристика (и сочетание характеристик) наблюдается такой же процент раз, как и в соответствующей генеральной совокупности.

C

Сдвиг или постоянный член для множественной регрессии (intercept or constant term, for multiple regression). Предсказанное значение Y, когда все X-переменные равны 0.

Сдвиг, a (intercept, a). Значение, которое линия регрессии отсекает на вертикальной оси; другими словами, значение Y, когда X равно 0.

Сезонная поправка (seasonal adjustment). Устранение из наблюдения ожидаемого сезонного компонента (путем деления соответствующего ряда на сезонный индекс для рассматриваемого периода), с тем чтобы один квартал или месяц можно было непосредственно сравнивать с другим, выявляя скрытые тенденции.

Сезонный индекс (seasonal index). Индекс для каждого периода года; указывает, насколько большими или меньшими значениями характеризуется данный период времени по сравнению с типичным для года периодом.

Сезонный компонент (seasonal component). В точности повторяющийся компонент временного ряда, который отражает тенденции, характерные для определенного периода года.

Систематическая выборка (systematic sample). Выборка, полученная путем выбора в основе выборки случайного начального места и последующим извлечением единиц, отделенных друг от друга фиксированным, регулярным интервалом. Несмотря на то что выборочное среднее, вычисленное для систематической выборки, представляет собой несмещенную оценку среднего значения генеральной совокупности (т.е. не является систематически завышенным или заниженным), этот метод порождает некоторые достаточно серьезные проблемы.

Скользящее среднее (moving average). Новый временной ряд, созданный путем усреднения расположенных рядом наблюдений.

Случайная выборка, или простая случайная выборка (random sample or simple random sample). Выборка, извлеченная таким образом, что (1) все единицы генеральной совокупности имеют равные вероятности быть отобранными и (2) все единицы отбираются независимо, без какого-либо взаимного влияния.

Случайная переменная (random variable). Определение или описание численного результата случайного эксперимента.

Случайное блуждание (random walk). Наблюдение чистого интегрированного (I) процесса, которое представляет собой случайный шаг от предыдущего наблюдения.

Случайные причины вариации (random causes of variation). Все те причины вариации, выявлять которые не имеет смысла.

Случайный эксперимент (random experiment). Любая хорошо определенная процедура, которая выдает наблюдаемый результат (исход), который невозможно точно предугадать заранее.

Смещенная выборка (biased sample). Выборка, которая с некоторой важной точки зрения не является репрезентативной.

Событие (event). Любая совокупность исходов (результатов), указанная заранее, до проведения случайного эксперимента; в результате каждого выполнения эксперимента событие либо наступает, либо не наступает.

Состояние статистического контроля [под контролем] (state of statistical control [in control]). Состояние процесса после того, как все неслучайные причины вариации удалось выяснить и устранить и остались лишь случайные причины.

Список чисел (list of numbers). Простейший вид совокупности данных, представляющий определенный тип информации (единственная статистическая переменная), измеренный для каждого исследуемого элемента (каждой элементарной единицы).

Среднее (mean). Общепринятый метод вычисления типичного значения для некоторой совокупности чисел путем сложения всех этих чисел и последующего деления полученной суммы на число элементов совокупности; используют также название *среднее значение* (average).

Среднее, или ожидаемое значение случайной переменной (mean or expected value of random variable). Типичное, или среднее значение случайной переменной.

Ссылка (reference). Примечание в отчете, указывающее на материал, заимствованный вами из стороннего источника; должно содержать информацию, достаточную для того, чтобы ваш читатель мог в случае необходимости самостоятельно обратиться к указанному источнику.

Стандартизованное значение (standardized number). Количество стандартных отклонений выше среднего (или ниже среднего, если стандартизованное значение отрицательно) определяется путем вычитания среднего и деления полученного результата на стандартное отклонение.

Стандартизованный коэффициент регрессии (standardized regression coefficient). Коэффициент $b_i S_{X_i} / S_Y$, представляющий ожидаемое изменение Y, вызванное изменением X_i; измеряется в единицах стандартного отклонения Y, приходящихся на одно стандартное отклонение X_i (в предположении, что все остальные X-переменные не изменяются).

Стандартная ошибка коэффициента наклона, S_b, для двумерных данных (standard error of the slope coefficient, S_b, for bivariate data). Указывает (приблизительно), насколько отстоит оценка наклона, b (коэффициент регрессии, вычисленный на данных выборки), от значения наклона в генеральной совокупности, β, вследствие случайности самой выборки.

Стандартная ошибка оценки, S_e (standard error of estimate, S_e). Приблизительная мера величины ошибок прогнозирования (остатков) для исследуемой совокупности данных; измерена в тех же единицах, что и Y.

Стандартная ошибка показателя (standard error of a statistic). Оценка стандартного отклонения выборочного распределения рассматриваемого показателя (статистики), приблизительно указывающая, насколько далеко от своего среднего значения (параметр генеральной совокупности) находится это значение статистики.

Стандартная ошибка прогноза (standard error for prediction). Мера неопределенности при прогнозировании, $S\sqrt{1+1/n}$; мера изменчивости расстояния между средним значением выборки и новым наблюдением.

Стандартная ошибка разности (standard error of the difference). Мера, необходимая при построении доверительных интервалов для разности средних и при выполнении проверки гипотез; эта мера дает оценку стандартного отклонения для разности выборочных средних.

Стандартная ошибка сдвига, S_a, для двумерных данных (standard error of the intercept term, S_a, for bivariate data). Указывает (приблизительно), насколько оценка a отстоит от α — истинной величины сдвига (отрезка, отсекаемого на вертикальной оси) в генеральной совокупности.

Стандартная ошибка среднего (standard error of the average). Указывает приблизительно величину разности между выборочным средним (случайным, наблюдаемым) $\bar{X}$ и средним значением генеральной совокупности (фиксированным, неизвестным) μ:

$$\text{Стандартная ошибка} = S_{\bar{x}} = S/\sqrt{n}\,.$$

Стандартное нормальное распределение (standard normal distribution). Нормальное распределение со средним $\mu = 0$ и стандартным отклонением $\sigma = 1$.

Стандартное отклонение [данные] (standard deviation [data]).Традиционный подход к измерению изменчивости; обобщает типичное расстояние между средним значением и отдельными значениями данных.

Стандартное отклонение [случайная переменная] (standard deviation [random variable]). Показатель риска, показывающий, насколько ожидаемое значение случайной переменной может отстоять от среднего.

Стандартное отклонение генеральной совокупности (population standard deviation). Обозначается символом σ и представляет собой меру изменчивости для всей генеральной совокупности.

Статистика (statistics). Наука и искусство сбора и анализа данных.

Статистика "хи-квадрат" (chi-squared statistic). Мера разности между фактическими частотами и ожидаемыми частотами (в предположении о справедливости нулевой гипотезы).

Статистически значимый (statistically significant). Результат, являющийся значимым на уровне 5% ($p < 0{,}05$). Используют также термины *высоко значимый* ($p < 0{,}01$), *очень высоко значимый* ($p < 0{,}001$) и *не значимый* ($p > 0{,}05$).

Статистический вывод (statistical inference). Процесс обобщения на основе данных выборки, позволяющий делать утверждения вероятностного характера относительно изучаемой генеральной совокупности.

Статистический показатель (статистика) (statistic). Какой-либо показатель, вычисленный на основе рассматриваемой выборки данных.

Статистическое управление качеством (statistical quality control). Использование статистических методов для оценивания и улучшения результатов какой-либо деятельности.

Статистическое управление процессом (statistical process control). Использование статистических методов для отслеживания функционирования некоторого процесса, позволяющее при необходимости вносить в этот процесс определенные коррективы или не вмешиваться в процесс, если он функционирует нормально.

Стационарный процесс (stationary process). Такие процессы, как процесс авторегрессии, процесс со скользящим средним и ARMA-модели, которые, как правило, ведут себя похожим образом на протяжении длительных периодов времени, оставаясь относительно близкими к своему долгосрочному среднему значению.

Степени свободы (degrees of freedom). Количество независимых источников информации в стандартной ошибке.

Стратифицированная случайная выборка (stratified random sample). Получается путем извлечения случайной выборки отдельно из каждой страты (из сегментов или групп) изучаемой генеральной совокупности.

Субъективная вероятность (subjective probability). Чье-либо мнение (лучше, если это будет мнение эксперта) о вероятности наступления события.

Т

Таблица вероятностей для стандартного нормального распределения (standard normal probability table). Таблица, содержащая значения вероятностей того, что случайная переменная со стандартным нормальным распределением окажется меньше, чем любое указанное значение.

Таблица случайных цифр (table of random digits). Перечень, в котором цифры от 0 до 9 появляются с вероятностью 1/10, независимо друг от друга.

Таблица совместных вероятностей (joint probability table). Таблица, содержащая вероятности двух событий, их дополнений и сочетания с использованием *и*.

Тенденция (тренд) для временного ряда (trend, for time series). *Очень* долгосрочное поведение исследуемого временного ряда; как правило, отображается в виде прямой линии или экспоненциальной кривой.

Теоретическая (идеальная) генеральная совокупность (idealized population). Достаточно большая, часто гипотетическая, генеральная совокупность, из которой извлечена и которую представляет имеющаяся выборка.

Теоретическая вероятность (theoretical probability). Значение, вычисленное с помощью точной формулы, основанной на определенной математической теории или модели (например, на правиле *равной вероятности*).

Тест наименьшего значимого различия для однофакторного дисперсионного анализа (least-significant-difference test, for one-way ANOVA). Используется лишь тогда, когда результат F-теста значим; тест сравнивает каждую пару выборок с целью выявления значимых отличий между ними.

Тест суммы рангов Вилкоксона (Wilcoxon rank-sum test). Способ вычисления результата непараметрического теста для двух независимых выборок.

Титульная страница (title page). Первая страница отчета, на которой указывается название отчета, фамилия и должность лица, для которого этот отчет составлен, ваша собственная фамилия и должность (как составителя отчета), а также дата.

У

Уравнение прогнозирования, или уравнение регрессии, для множественной регрессии (prediction equation or regression equation, for multiple regression). Вычисленное $Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$, которое может использоваться для прогнозирования или управления.

Уровень тестирования, или уровень значимости (test level or significance level). Вероятность неправильно принять альтернативную гипотезу, когда на самом деле правильной является нулевая гипотеза (т.е. вероятность совершить ошибку первого рода). Как правило, используют уровень 5%, хотя иногда используют уровни 1% или 0,1% (или даже — в некоторых областях исследований — 10%), для чего нужно выбрать соответствующий столбец t-таблицы.

Условная вероятность (conditional probability). Вероятность наступления события, включающая информацию о том, что некоторое другое событие уже произошло (вероятность наступления события A *при условии*, что уже наступило событие B).

Условные проценты генеральной совокупности для двух качественных переменных (conditional population percentages for two qualitative variables). Вероятности категорий одной переменной при условии, что рассмотрение ограничивается только одной категорией второй переменной.

Уточненная стандартная ошибка (adjusted standard error). Используется в случае, когда размер генеральной совокупности невелик и выборка является существенной частью генеральной совокупности. Вычисляется путем введения в формулу стандартной ошибки поправочного коэффициента, учитывающего конечный объем генеральной совокупности:

(коэффициент поправки на конечность генеральной совокупности) ×
× (стандартная ошибка) =

$$= \sqrt{\frac{N-n}{N}} S_{\bar{x}} = \sqrt{\frac{N-n}{N}} \frac{S}{\sqrt{n}}.$$

Ф

Функция кумулятивного распределения (cumulative distribution function). Предназначенный для отображения перцентилей график, на котором значения процентов показаны как функция от значений данных.

Функция оценки (estimator). Статистика, которая вычисляется на основе данных выборки и используется в качестве предполагаемого значения параметра генеральной совокупности.

Х

"Хи-квадрат" тест независимости (chi-squared test for independence). Проверка независимости двух качественных переменных, основанная на таблице наблюдаемых частот из двумерной совокупности качественных данных.

"Хи-квадрат" тест равенства процентов (chi-squared test for equality of percentages). Проверка, которая используется для определения, действительно ли источником таблицы наблюдаемых частот или процентов (описывающей некоторую качественную переменную) может быть генеральная совокупность с известными значениями процентов (которые выступают в качестве эталонных значений).

"Хи-квадрат" тесты (chi-squared tests). Тесты, которые используются при проверках статистических гипотез для качественных данных, когда мы имеем дело не с числами, а с категориями.

Ц

Центральная предельная теорема (central limit theorem). Правило, утверждающее, что для случайной выборки объемом n наблюдений из некоторой генеральной совокупности (1) выборочное распределение среднего и суммы при увеличении n все

больше приближается к нормальному, (2) средние значения и стандартные отклонения распределений среднего и суммы имеют следующий вид (где μ — среднее отдельных значений, σ — стандартное отклонение этих отдельных значений):

$$\mu_{\bar{x}} = \mu; \qquad \mu_{sum} = n\mu;$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}; \qquad \sigma_{sum} = \sigma\sqrt{n}.$$

Циклический компонент (cyclic component). Среднесрочный компонент временного ряда, состоящий из последовательных подъемов и снижений, которые не повторяются каждый год.

Ч

Частота ложной тревоги для управления качеством (false alarm rate, for quality control). Частота принятия решений о необходимости вмешательства в процесс, когда такое вмешательство не требуется.

Частотный (не байесовский) анализ (frequentist [non-Bayesian] analysis). Анализ, в котором для соответствующих вычислений не используются субъективные вероятности, хотя и его все же нельзя назвать полностью объективным, поскольку субъективные мнения влияют на отбор данных и выбор модели (математической основы).

Числовая ось (number line). Прямая линия (обычно горизонтальная), значения шкалы на которой указываются числами, записанными под этой линией.

Чистый интегрированный (I) процесс (pure integrated [I] process). Процесс, в котором каждое следующее наблюдение представляет собой случайный шаг в направлении от текущего наблюдения; также называется *случайным блужданием* (random walk).

Э

Экономная модель для временных рядов (parsimonious model, for time series). Модель, в которой для описания сложного поведения временного ряда используются лишь оценки нескольких параметров.

Экспоненциальное распределение (exponential distribution). Чрезвычайно асимметричное непрерывное распределение, которое может описывать, например, время ожидания или продолжительность телефонных разговоров.

Экстраполяция (extrapolation). Прогнозирование за пределами диапазона данных, имеющихся в распоряжении исследователя; связано с особым риском, поскольку в таком случае результат нельзя проверить с помощью изучения имеющихся данных.

Эластичность Y по отношению к X_i (elasticity Y with respect to X_i). Ожидаемое процентное изменение Y, связанное с увеличением X_i на 1% при условии, что все другие X-переменные не изменяются; оценивается с помощью коэффициента регрессии, полученного в результате регрессионного анализа логарифмов Y и X_i.

Элементарные единицы (elementary units). Отдельные предметы или вещи (например, люди, домашние хозяйства, фирмы, города, телевизоры и т.п.), результаты измерения свойств которых составляют определенную совокупность данных.

Эталонное значение (reference value). Заданное, конкретное число, полученное не из данных той выборки, со средним значением которой это число сравнивается; обозначается μ_0.

Эффективный тест (effective test). Тест, который использует содержащуюся в данных информацию более эффективно, чем какой-либо другой тест.

Предметный указатель

A

ARIMA-процессы Бокса–Дженкинса, *771*

D

Dow Jones Industrial, индекс фондовой биржи, *180*

E

Критерий
- знаков, *850*
- разности средних рангов, *858*
- суммы рангов Вилкоксона, *858*

F

F-статистика, *809*
F-таблица, *817*
F-тест, *809*

П

Процесс
- авторегрессии и скользящего среднего (ARMA), *778*
- авторегрессии, *775*
- авторегрессионного интегрированного скользящего среднего (ARIMA), *785*
- скользящего среднего, *776*
- случайного шума, *774*

T

t-тест
- (t-критерий) Стьюдента, *459*
- t-статистика, *459*
- две зависимые выборки, *480*
- две независимые выборки, *482*
- критическое значение, *459*
- односторонний, *469*

U

U-критерий Манна–Уитни, *858*

А

Анализ трендов и сезонности, *755*

Б

Бимодальное распределение, *88*
Блочная диаграмма, *139*

В

Вероятность, *28*, *36*
Вероятность события, *229*
Взвешенное среднее, *121*
Внутривыборочная вариация, *813*
Внутригрупповая вариация, *813*
Временной ряд, *50*
Временные ряды
- математическая модель, процесс, *744*
- нерегулярный компонент, *755*
- сезонный компонент, *755*
- тренд, *755*
- циклический компонент, *755*

Вторичные данные, *52*
Выборка, *32*, *341*
- без возврата, *344*
- репрезентативность, *32*
- с возвратом, *344*
- случайная, *32*

Выборочное пространство, *224*
Выборочное распределение, *352*
Выборочный параметр, *345*
Выброс, *139*, *542*
Выбросы (сильно отклоняющиеся значения), *91*

Г

Генеральная совокупность, *32*, *341*
Гетероскедастичность, *537*
Гипотеза, *446*
- альтернативная, *447*
- исследовательская, *447*
- нулевая, *446*

Гистограмма, *70*, *73*, *77*

Д

Данные, *29*
Данные об одном временном срезе, *50*
Двумерные данные, *44*
Двусторонняя проверка
статистической гипотезы о среднем, *450*
Дерево
вероятностей, *249*
решений, *249*
Диагностическая диаграмма, *662*
Диаграмма Венна, *237*
Диаграмма Парето, *913*
Диаграмма рассеяния, *519*, *520*
Диверсификация, *187*
Дискретная переменная, *47*
Дискретная случайная величина, *279*
Дисперсия, *170*
вычисление для выборки, *174*
Дневная прибыль, *180*
Доверительная вероятность, *397*
Доверительный
интервал, *33*, *397*
уровень, *397*
Дополнение события, *238*
Доу Джонс, индекс фондовой биржи, *180*

З

Зависимые события, *246*
Заданное (опорное) значение в проверке статистических гипотез, *450*
Закон больших чисел, *231*

И

Изменчивость, *170*
Индикаторная переменная, *684*
Интервал предсказания, *423*
Исследование, *31*

К

Карта контроля, *915*
R-карта, *919*
X-карта, *919*
контрольные границы, *915*
процентная карта, *926*
центральная линия, *915*
Качественные данные, *49*
Квартили, *137*
Кластеринг, *540*
Ковариация, *524*
Количественные данные, *47*
Контроль качества
неслучайная причина отклонений, *912*
случайная причина отклонений, *912*
уровень ложной тревоги, *916*
Корреляционная матрица, *651*
Корреляция, *519*, *521*
Коэффициент вариации, *170*, *192*, *674*
Коэффициент доверия, *397*
Коэффициент корреляции, *521*
Критерий
"хи-квадрат", *879*, *881*
независимость двух качественных переменных, *889*
Кумулятивный процент, *913*

Л

Линейная взаимосвязь, *525*
Линия наименьших квадратов, *549*
Логарифмирование данных, *86*
Ложная корреляция, *545*

М

Медиана, *126*
Межвыборочная вариация, *813*
Межгрупповая вариация, *813*
Метод "ствол и листья", *97*
Метод Байеса, *236*
Многомерные данные, *45*
Множественная регрессия, *611*
Мода, *133*
Модель множественной линейной регрессии, *626*

Н

Наблюдаемое значение, *279*
Наблюдение, *613*
Набор данных, *42*
Наклон, *549*
Независимость двух качественных переменных, *888*
Независимые события, *246*
Нелинейная взаимосвязь, *533*
Неопределенность, *170*
Непараметрические методы, *847*
Непрерывная переменная, *47*
Непрерывная случайная величина, *279*
Неравная вариация, *537*
Несимметричное (скошенное) распределение, *81*
Несмещенная оценка, *345*
Несовместимые события, *240*
Нестационарный процесс, *782*
Номинальные данные, *50*
Нормальное распределение, *79*, *295*
Нормированное значение, *303*

О

Обобщение, *117*
Объединение событий, *240*
Одномерные данные, *43*
Односторонний доверительный интервал, *419*
Однофакторный дисперсионный анализ, *809*
F-статистика, *815*, *817*
F-тест, *812*
внутригрупповая вариация, *816*
межгрупповая вариация, *815*
общее (главное) среднее, *815*
проверка наименьшего значимого различия, *824*
Опцион, *534*
Основа генеральной совокупности, *343*
Отклонение, *172*
Относительная частота события, *230*
Оценка, *31*, *345*
неизвестной величины, *33*
Оценочная функция параметра, *345*
Ошибка
I рода, *462*
II рода, *463*
Ошибка оценки, *345*

П

Параметр
выборки, *345*
генеральной совокупности, *345*
Параметрические методы, *848*
Первичные данные, *52*
Переменная, *42*
Переменные затраты, *45*
Переменные издержки, *196*
Перепись, *345*
Пересечение событий, *239*
Перцентили, *137*
Планирование, *31*
исследования, *32*
Подробная блочная диаграмма, *139*
Полиномиальная регрессия, *679*
Поправка на сезонные колебания, *763*
Порядковые (ординальные) данные, *49*
Постоянные затраты, *45*
Правило равной вероятности, *233*
Предварительное исследование данных, *32*
Преобразование данных, *85*
Причинность, *544*
Пробное (пилотное) исследование, *352*
Проверка гипотезы, *31*, *34*
биномиальный случай, *457*
доверительная вероятность (p-значение), *466*
о среднем значении, *451*
статистическая значимость, *465*
статистической, *446*
уровень значимости, *466*
условия применимости, *464*
Произведение событий, *239*
Простая случайная выборка, *346*
Противоположное событие, *238*
Процесс, *911*
Пять базовых показателей, *138*

Р

Разброс, *170*
Размах, *170*, *189*
Ранг, *126*
Распределение Пуассона, *315*
Рассеяние, *170*
Регрессионный анализ, *546*

линейный, *547*
Регрессия
F-тест, *614*, *628*
автоматизированный выбор переменных, *658*
взаимодействие переменных, *682*
коэффициент, *549*, *614*
детерминации, *558*, *625*
стандартизованный, *645*
мультиколлинеарность, *649*, *650*
остаток, *551*, *614*
постоянный член, *549*
сдвиг, *614*
стандартная ошибка
коэффициента регрессии, *560*
оценки, *555*
сдвига, *561*
Результат случайного эксперимента, *226*
Репрезентативность выборки, *343*
Риск, *281*

С

Сдвиг, *549*
Сезонная поправка, *756*
Сезонный индекс, *755*, *760*
Систематическая выборка, *372*
Скользящее среднее, *756*, *757*
Скорректированная стандартная ошибка, *363*
Случайная величина, *279*
Случайная выборка, *346*
Случайное блуждание, *533*, *782*
Случайный эксперимент, *223*
Смещение выборки, *343*
Событие, *227*
Среднее, *118*
квадратическое отклонение, *170*
квадратичное отклонение, *170*
Среднее дискретной случайной величины, *281*
Стандартная ошибка
предсказания, *423*
среднего, *360*
статистики, *359*
Стандартное
нормальное распределение, *298*
отклонение дискретной случайной величины, *281*
Стандартное отклонение, *170*, *171*, *188*
вычисление для выборки, *173*
для генеральной совокупности, формула, *188*
интерпретация, *175*
Статистика, *29*, *345*
"хи-квадрат", *879*
ожидаемая частота, *879*
фактическая частота, *879*
Статистический
анализ, *28*
вывод, *32*, *397*
контроль
качества, 908
процесса, 911
Стационарный процесс, *782*
Степени свободы, *403*
Столбиковая диаграмма, *77*
Стратифицированная случайная выборка, *368*
Субъективная оценка вероятности, *234*
Сумма событий, *240*

Т

Таблица вероятностей для стандартного нормального распределения, *298*
Таблица случайных чисел, *347*
Таблица совместных вероятностей, *257*
Таблица частот или процентов, *880*
Теоретическая генеральная совокупность, *364*
Теоретическое значение вероятности, *233*
Теория эффективного рынка, *532*
Типическое значение, *118*

У

Условная вероятность, *243*

Ф

Фиксированные затраты, *196*
Фиктивная переменная, *684*
Форма типового отчета, *722*
Функция кумулятивного распределения, *142*

Х

Хеджирование, *613*

Ц

Центральная предельная теорема, *353*

Ч

Частотный анализ, *236*
Числовая ось, *72*
Чистый интегрированный процесс, *782*

Ш

Шанс, *230*

Э

Экспоненциальное распределение, *318*
Экстремумы, *137*
Эластичность, *674*
Элементарная единица анализа, *42*
Эффективность статистического теста, *848*
Эффективный финансовый рынок, *248*

Научно-популярное издание

Эндрю Сигел

Практическая бизнес-статистика

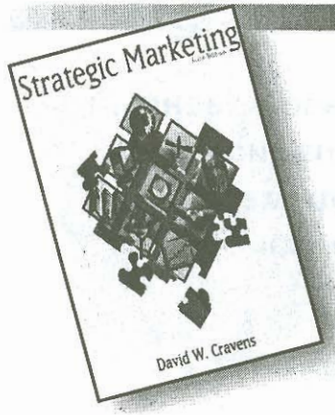
Литературные редакторы *И.А. Попова, И.А. Шишкина*
Верстка *В.В. Терещенко*
Художественные редакторы *А.В. Говдя, В.Г. Павлютин*
Корректоры *Л.А. Гордиенко, Т.А. Корзун, Л.В. Коровкина, О.В. Мишутина, Л.В. Чернокозинская*

Издательский дом "Вильямс".
101509, Москва, ул. Лесная, д. 43, стр. 1.
Изд. лиц. ЛР № 090230 от 23.06.99
Госкомитета РФ по печати.

Подписано в печать 29.07.2002. Формат 70×100/16.
Гарнитура Times. Печать офсетная.
Усл. печ. л. 79,11. Уч.-изд. л. 63,29.
Тираж 3000 экз. Заказ № 1029.

Отпечатано с диапозитивов в ФГУП "Печатный двор"
Министерства РФ по делам печати,
телерадиовещания и средств массовых коммуникаций.
197110, Санкт-Петербург, Чкаловский пр., 15.

Дэвид В. Кревенс



Стратегический маркетинг, 6-е издание

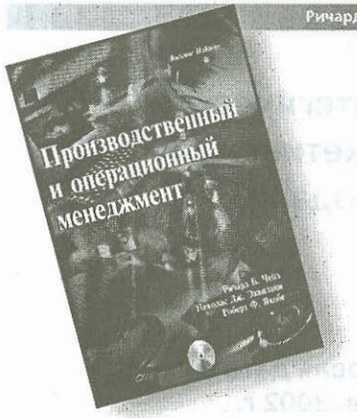
Плановая дата выхода – 4-й кв. 2002 г.

В современной рыночной экономике происходит повсеместное ужесточение конкурентной борьбы. Эта книга, благодаря удачному сочетанию теоретического и практического материала, подскажет, как разрабатывать такие стратегии маркетинга, которые позволят вашей компании удержаться на плаву в бушующем море конкуренции, кроме этого, вы узнаете как найти и развить уникальные конкурентные преимущества, которые обеспечат вашему бизнесу устойчивую конкурентную позицию.

Особое внимание автор этой книги уделил таким стратегически важным проблемам организации маркетинговой деятельности на предприятии, как рыночная ориентация всех функциональных подразделений компании; установление и поддержание долгосрочных отношений с потребителями, поставщиками, посредниками; массовая индивидуализация выпускаемой продукции; использование бенчмаркинга для повышения уровня конкурентоспособности товаров и услуг; использование новых организационных форм для повышения эффективности маркетинговых решений.

Книга предназначена для студентов старших курсов, менеджеров различного уровня.

Ричард Б. Чейз, Николас Дж. Эквилайн, Роберт Ф. Якобс



Производственный и операционный менеджмент, 5-е издание

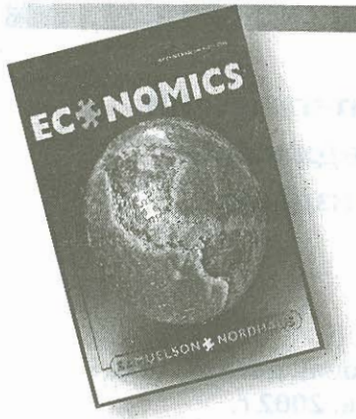
В продаже

В книге рассмотрен весь спектр вопросов операционного менеджмента — от выбора бизнес-стратегии и проектирования продукта или услуги до полного обновления продуктовой, производственной и организационной базы. Приведены и проанализированы новейшие идеи в управлении производством и сферой предоставления услуг.

Авторы построили изложение материала таким образом, чтобы как можно точнее описать этапы жизненного цикла производственной системы. Помимо очень детального изложения теории, авторы уделили достаточно внимания описанию практических приемов и методик. Для закрепления материала, изложенного в книге, читателям предложено большое количество задач и ситуаций для анализа. К части из них дается подробное описание алгоритма решения или анализа.

Книга написана простым для понимания языком, с минимальным применением специальной терминологии и математических выкладок, анализирует много практических ситуаций, прекрасно иллюстрирована примерами в тексте и на прилагаемом компакт-диске и потому доступна самому широкому кругу читателей.

П. Самуэльсон и В. Нордхаус

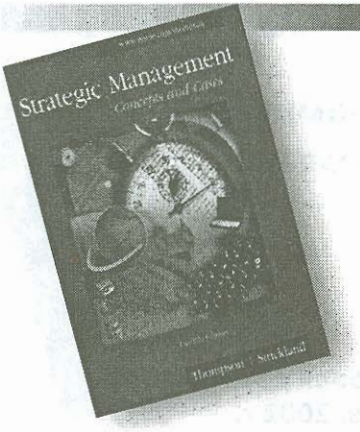


Экономика, 17-е издание

Плановая дата выхода – 4-й кв. 2002 г.

Экономическая теория —многоплановая наука. Это, с одной стороны упрощает, а с другой, — усложняет задачу написания учебника по этому предмету. По общему мнению книга П. Самуэльсона и В. Нордхауса стала классическим изданием по теории экономики. Эта книга тщательно продумана, чтобы быть полезной для изучающих экономическую теорию в любом объеме. Более серьезные материалы собраны в дополнительных приложениях или специально обозначенных разделах. Эти материалы предназначены для наиболее любознательных студентов, а также студентов, которым требуется основательно изучить весь курс. Обучаясь по упрощенной программе, сложные разделы можно опустить, изучая только суть экономического анализа и не теряя при этом нити экономических рассуждений. Эта книга является вызовом и самым упорным молодым ученым. Действительно, многие из ведущих экономистов, наших современников, говорят, что книга Экономика всегда была их “путеводной звездой” на пути к вершинам научной карьеры.

В новом издании сохранены главные достоинства всех предшествующих: простота изложения, четкая структура, богатый иллюстративный материал. Однако в этот раз авторы уделили значительно большее внимание глобальным проблемам экономического развития, влиянию новых технологий на этот процесс, а также новым тенденциям развития экономической мысли и проблемам стран с переходной экономикой. Существенно обновлен и статистический материал.



Стратегический менеджмент, 12-е издание

Плановая дата выхода – 4-й кв. 2002 г.

Стратегический менеджмент – классическое учебное пособие А. А. Томпсона и А. Дж. Стрикленда-младшего, неоднократно переиздававшееся, но не теряющее популярности в силу актуальности излагаемого материала. Сегодня, в эпоху глобализации экономики, ни одна компания не может существовать без четко сформулированной стратегии развития. Теперь даже малый бизнес рискует быть уничтоженным конкурентами, если не осознает своего положения в отрасли и не определит своих долгосрочных перспектив с учетом действий соперников, не только ближайших, но и находящихся на другой стороне планеты.

Предлагаемое издание книги *Стратегический менеджмент* состоит из двух частей. Первая, теоретическая, знакомит читателя с основными положениями, концепциями и понятиями. В ней дан обзор современного бизнеса, значительно изменившегося с появлением электронных технологий. Не случайно целую главу авторы посвящают разработке бизнес-моделей и стратегий для Internet-экономики.

Во второй, значительно превосходящей первую по объему, части представлены ситуации для анализа – снабженные таблицами и диаграммами истории различных компаний, крупных и маленьких, находящихся на гребне успеха или на краю гибели. Разнообразие примеров позволяет каждому заинтересованному читателю найти аналогии с собственным бизнесом и почерпнуть нечто полезное.