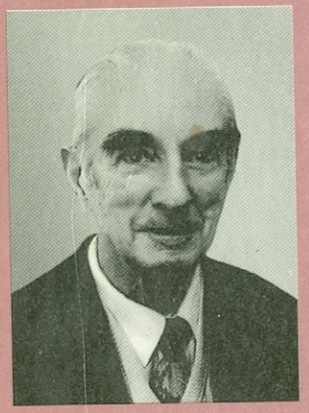


Елисеева Ирина Ильинична –
доктор экономических наук, профессор, член-корреспондент Российской академии наук, член Международного статистического института, Европейского общества истории экономической мысли.
Автор книг "Статистические методы измерения связей" (1982), "Структура и функции семейных групп" (1983), "История статистики"(1990), "Моя профессия – статистик" (1991) и др.; соавтор и руководитель авторских коллективов учебников "Социальная статистика" (1-е изд. – 1998), "Эконометрика" (2000), "Статистика" (2001) и др.
Научный редактор переводов на русский язык фундаментальных работ в области финансового учета и анализа, рынка ценных бумаг, экономической теории, методов принятия решений.
Область научных интересов: социальная статистика, история статистики.



Юзбашев Михаил Михайлович –
доктор экономических наук, профессор.
Автор книг "Методы изучения динамики распределений и зависимостей"(1972), "Статистический анализ тенденций и колеблемости"(1983), "Анализ временных рядов и прогнозирование"(2001) и др.
Область научных интересов: методы анализа временных рядов; применение статистических методов в сельском хозяйстве.

ISBN 5-279-02414-7

И.И.Елисеева
М.М.Юзбашев

ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ

$$\chi^2 = \sum \frac{(m-m')^2}{m'}$$

ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ

И.И.Елисеева
М.М.Юзбашев

$$R = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\delta_x \delta_y}; \quad E(x$$

И.И.Елисеева, М.М.Юзбашев

ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ

Под редакцией члена-корреспондента
Российской Академии наук
И.И.Елисеевой

ПЯТОЕ ИЗДАНИЕ,
ПЕРЕРАБОТАННОЕ И ДОПОЛНЕННОЕ

Рекомендовано
Министерством образования
Российской Федерации
в качестве учебника
для студентов высших учебных заведений,
обучающихся по направлению и специальности
"Статистика"

2710

Москва
"Финансы и статистика"
2004

.. 9
.. 13
.. 13
и-
.. 17
.. 23
о-
.. 27
.. 30
.. 31
.. 32
.. 32
.. 38
.. 50
.. 57
.. 64
е-
.. 72
.. 75
.. 79
.. 81

УДК 311(075.8)
ББК 60.6я73
Е51

РЕЦЕНЗЕНТЫ:

Кафедра общей теории статистики
Московского государственного университета
экономики, статистики и информатики (МЭСИ);

Г.Л. Громыко,
доктор экономических наук,
профессор кафедры статистики МГУ им. М.В. Ломоносова

Елисеева И.И., Юзбашев М.М.

Е51 Общая теория статистики: Учебник / Под ред. И.И. Елисеевой. — 5-е изд., перераб. и доп. — М.: Финансы и статистика, 2004. — 656 с.: ил.

ISBN 5-279-02414-7

Рассматриваются вопросы организации статистики, ее особенности в нашей стране и других странах. Пятое издание (4-е изд. — 1999 г.) расширено изложением методов сбора данных с использованием современных технологий, приведены методы многомерной классификации, раскрыта основополагающая роль вариационного анализа. Описание теоретических основ выборочного метода и методов проверки статистических гипотез сочетается с примерами их использования в исследованиях и в работе органов государственной статистики. Особое внимание уделено статистическому анализу неколичественных переменных систем регрессионных уравнений, анализу временных рядов.

Для студентов высших учебных заведений, обучающихся по направлению и специальности «Статистика».

Е $\frac{0702000000-227}{010(01)-2004}$ 269–2003

УДК 311(075.8)
ББК 60.6я73

ISBN 5-279-02414-7

© И. И. Елисеева, М. М. Юзбашев, 1995
© И. И. Елисеева, М. М. Юзбашев, 2004

ОГЛАВЛЕНИЕ

Предисловие 9

Глава 1. Понятие о статистике 13
1.1. Что такое статистика 13
1.2. Статистическая закономерность. Статистические совокупности 17
1.3. Признаки и их классификация 23
1.4. Определение предмета статистики — основа статистической методологии 27
Резюме 30
Рекомендуемая литература 31

Глава 2. Организация статистики. Статистическое наблюдение 32
2.1. Организация государственной статистики в Российской Федерации 32
2.2. Важнейшие международные организации и их статистические службы 38
2.3. Требования, предъявляемые к собираемым данным. Формы организации и виды статистического наблюдения 50
2.4. Подготовка статистического наблюдения . . 57
2.5. Статистическая отчетность 64
2.6. Ошибки статистического наблюдения. Методы контроля данных наблюдения 72
2.7. Реформирование российской государственной статистики 75
Резюме 79
Рекомендуемая литература 81

Глава 3. Статистические показатели 82
3.1. Сущность и значение статистических показателей. Показатель и его атрибуты 82
3.2. Классификация статистических показателей 85
3.3. Общие принципы построения относительных статистических показателей. 91
3.4. Понятие о системах статистических показателей 93
3.5. Функции статистических показателей. 95
Резюме 98
Рекомендуемая литература 99

Глава 4. Представление статистических данных: таблицы и графики 100
4.1. Статистические таблицы 100
4.2. Основные виды графиков. 106
4.3. Картограммы и картодиаграммы 115
Резюме 119
Рекомендуемая литература 119

Глава 5. Средние величины и изучение вариации 120
5.1. Однородность и вариация массовых явлений 120
5.2. Средняя арифметическая величина 123
5.3. Другие формы средних величин 134
5.4. Средняя величина как выражение закономерности 138
5.5. Вариация массовых явлений. 140
5.6. Построение вариационного ряда. Виды рядов. Ранжирование данных 142
5.7. Структурные характеристики вариационного ряда 150
5.8. Показатели размера и интенсивности вариации 154
5.9. Моменты распределения и показатели его формы 160
5.10. Предельно возможные значения показателей вариации и их применение 165
Резюме 169
Рекомендуемая литература 171

Глава 6. Группировка 172
6.1. Значение и сущность группировки 172
6.2. Виды группировок 177
6.3. Многомерные группировки 191
Резюме 212
Рекомендуемая литература 213

Глава 7. Выборочное наблюдение. Испытание статистических гипотез 214
7.1. Причины применения выборочного наблюдения. Дескриптивная статистика и статистический вывод 214
7.2. Способы отбора, обеспечивающие репрезентативность выборки. Виды выборки . . . 218
7.3. Ошибка выборки 223
7.4. Влияние вида выборки на величину ошибки выборки 231
7.5. Задачи, решаемые при применении выборочного метода. 240
7.6. Распространение данных выборочного наблюдения на генеральную совокупность . 247
7.7. Малая выборка. 250
7.8. Примеры применения выборочного метода 253
Резюме 267
Рекомендуемая литература 269

Глава 8. Статистическая проверка гипотез 270
8.1. Общие понятия 270
8.2. Проверка гипотезы о законе распределения 274
8.3. Проверка гипотезы о связи на основе критерия χ^2 (хи-квадрат) 287
8.4. Проверка гипотезы о средних величинах . . 292
8.5. Основы дисперсионного анализа 295
8.6. Некоторые непараметрические критерии . . 305
Резюме 317
Рекомендуемая литература 319

Глава 9. Корреляционно-регрессионный анализ и моделирование статистических связей . . . 320
9.1. Понятие о статистической и корреляционной связи . . . 320
9.2. Условия применения и ограничения корреляционно-регрессионного метода . . . 324
9.3. Задачи корреляционно-регрессионного анализа и моделирования . . . 327
9.4. Вычисление и интерпретация параметров парной линейной регрессии . . . 334
9.5. Статистическая оценка надежности параметров парной регрессии и корреляции . . . 345
9.6. Применение линейного уравнения парной регрессии . . . 349
9.7. Вычисление параметров парной линейной регрессии на основе аналитической группировки . . . 352
9.8. Параболическая корреляция . . . 358
9.9. Гиперболическая корреляция . . . 361
9.10. Множественное уравнение регрессии . . . 364
9.11. Меры тесноты связей в многофакторной системе . . . 370
9.12. Вероятностные оценки параметров множественной регрессии и корреляции . . . 380
9.13. Корреляционно-регрессионные модели и их применение в анализе и прогнозе . . . 382
Резюме . . . 389
Рекомендуемая литература . . . 391

Глава 10. Системы регрессионных уравнений . . . 392
10.1. Понятие о системах регрессионных уравнений . . . 392
10.2. Проблемы решения систем взаимосвязанных уравнений . . . 394
10.3. Преобразование структурных уравнений в приведенные и их идентификация . . . 397
10.4. Косвенный метод наименьших квадратов 401
10.5. Двойной метод наименьших квадратов . . . 405
Резюме . . . 409
Рекомендуемая литература . . . 410

Глава 11. Статистический анализ неколичественных переменных . . . 411
11.1. Зависимость методов измерений связей от уровня измерения переменных . . . 411
11.2. Измерение связи между двумя дихотомическими переменными . . . 416
11.3. Измерение связи по таблицам взаимной сопряженности $m \times p$ . . . 422
11.4. Теоретико-информационные меры связей 429
11.5. Другие меры связей между номинальными переменными . . . 434
11.6. Коэффициенты корреляции рангов . . . 437
11.7. Коэффициент конкордации . . . 441
Резюме . . . 443
Рекомендуемая литература . . . 444

Глава 12. Статистическое изучение динамики . . . 445
12.1. Виды динамических рядов. Сопоставимость данных в изучении динамики . . . 445
12.2. Элементы динамики: основная тенденция и колебания . . . 447
12.3. Показатели, характеризующие тенденцию динамики . . . 449
12.4. Особенности показателей динамики для рядов, состоящих из относительных уровней . . . 455
12.5. Средние показатели тенденции динамики 459
12.6. Методы выявления типа тенденции динамики . . . 468
12.7. Методика измерения параметров тренда . . . 476
12.8. Методика изучения и показатели колеблемости . . . 486
12.9. Измерение устойчивости в динамике . . . 493
12.10. Сезонные колебания и полное разложение дисперсии уровней динамического ряда . . . 496
12.11. Прогнозирование на основе тренда и колеблемости . . . 511
12.12. Корреляция рядов динамики . . . 516
Резюме . . . 522
Рекомендуемая литература . . . 525

Глава 13. Индексы 526
13.1. Понятие индекса 526
13.2. Индекс как показатель центральной тенденции (индекс средний из индивидуальных) 528
13.3. Агрегатные индексы. Система индексов . . . 537
13.4. Свойства индексов 545
13.5. Индексный анализ взвешенной средней. Индекс структуры 548
13.6. Построение индексов при обобщении данных по единицам совокупности и по элементам 552
13.7. Границы и условия применения индексного метода 563
13.8. Комплексное использование индексного и регрессионного методов анализа 568
13.9. Примеры использования индексов в экономико-статистических расчетах 585
Резюме 594
Рекомендуемая литература 596

Глава 14. Статистическое изучение структуры совокупности и ее изменений 597
14.1. Показатели простой (одномерной) структуры 597
14.2. Показатели иерархической (древовидной) структуры 599
14.3. Показатели балансовой структуры 603
14.4. Показатели многомерной структуры с пересекающимися признаками 609
14.5. Сравнительный анализ структур 611
14.6. Показатели концентрации, специализации, монополизации. Многомерная структура . . 617
14.7. Абсолютные и относительные показатели изменения структуры 621
14.8. Ранговые показатели изменения структуры 625
Резюме 628
Рекомендуемая литература 629

Приложения. 1. Статистико-математические таблицы 630
2. Основные принципы официальной статистики в регионе Европейской экономической комиссии 647
Предметный указатель 649

ПРЕДИСЛОВИЕ

«Общая теория статистики» — одна из основных дисциплин в системе экономического образования и важнейшая для тех, кто избрал статистику своей профессией.

Термин «статистика» возник во второй половине XVIII в. в связи *с познанием государств*, как тогда предпочитали говорить, описанием их особенностей, достопримечательностей. К тому же времени относится начало преподавания статистики в университетах Германии.

История человечества показала, что без статистических данных невозможны управление государством, развитие отдельных отраслей и секторов экономики, обеспечение оптимальных пропорций между ними. Необходимость сбора и обобщения множества данных о населении страны, предприятиях, банках, фермерских хозяйствах и т.д. привела к возникновению специальных статистических служб — учреждений государственной статистики. В зависимости от отрасли, по которой организуются измерения, сбор, обработка и анализ статистических данных, различают статистику населения, промышленности, сельского хозяйства, капитального строительства, финансов и др. Все эти разделы статистики призваны вырабатывать методы статистической работы для отражения процессов в соответствующей отрасли. Рассчитываются статистические показатели и для экономики в целом — валовой национальный продукт, валовой внутренний продукт, совокупная добавленная стоимость, уровень инфляции и т.д.

Статистик нужен и для страны, и для предприятия, и для региона. Статистические методы позволяют разрабатывать стратегию развития фирмы на основе прогнозирования динамики основных показателей и соотношений между ними. Важное значение для успешной работы фирмы имеют статистические методы контроля и анализа качества продукции. Динамика макроэкономических показателей дает основание для разработки перспективных планов развития экономики в

целом, измерения эффективности общественного производства и т.д.

Несмотря на разнообразие сфер применения статистики, имеются общие методы статистической работы, которыми нужно руководствоваться всегда и везде. Именно с такими правилами сбора, обработки и анализа статистических данных знакомит курс «Общая теория статистики».

Статистик работает с числовой и нечисловой информацией, с большими и малыми выборками, с вычислениями, таблицами и графиками. Имеется множество отечественных и зарубежных пакетов прикладных программ статистической обработки данных на персональных компьютерах.

Разработаны специальные, предназначенные для обучения студентов программы, которые содержат подробные объяснения статистических методов и тесты для проверки знаний.

С развитием рыночной экономики — увеличением числа хозяйственных единиц, их типов, развитием аудита, финансового менеджмента задачи отечественной статистики значительно расширились. В практику внедряются методики, принятые в международной статистике. Возрастают потребности в статистическом моделировании и прогнозировании, в использовании выборочных оценок.

В учебнике рассмотрены основные процедуры сбора, обработки и анализа массовых данных, возможности их реализации на персональных компьютерах. Особое внимание уделено обоснованию вероятностного характера статистического вывода, выборочному методу, проверке статистических гипотез.

Этот учебник дает представление об основных статистических методах, их возможностях и границах применения. Для желающих более глубоко изучить соответствующий раздел статистики в конце каждой главы приведен список рекомендуемой литературы.

Авторы стремились показать, что статистика не является скучной и трудной наукой, как иногда думают, а ее изучение может доставить удовольствие. Этим обусловлена подача материала — неформальная, но информативная.

Изложение теории проиллюстрировано примерами из разнообразных областей, которые должны убедить читателя во «всесильности» статистики, возможности ее применения при решении различных задач.

Учебник соответствует программе подготовки бакалавров. Вместе с тем он будет полезен и занимающимся в магистратуре и даже в аспирантуре. В данное, 5-е издание, внесены уточнения и дополнения во все главы. Глава 2 существенно переработана и дополнена с учетом изменений в работе государственной статистики. Выборочный метод излагается теперь отдельно от методов проверки статистических гипотез, дополненных прежде всего изложением непараметрического тестирования.

Каждая глава завершается резюме, подытоживающим основные положения. Это должно способствовать лучшему усвоению материала.

Заново написаны главы 10 и 11, посвященные соответственно системам уравнений регрессии и статистическому анализу нечисловой информации. Материал глав 9—12 можно с успехом использовать при изучении дисциплины «Эконометрика».

В приложении увеличено количество статистико-математических таблиц: в дополнение к традиционному набору, включающему таблицу вероятностей нормального распределения, плотности вероятностей нормального распределения, распределения Стьюдента, таблицу критических значений статистики хи-квадрат, F-критерия, таблицу случайных чисел, критических значений коэффициента корреляции, даны таблицы критических значений критерия знаков Вилкоксона, критических значений двухвыборочного критерия рангов Вилкоксона, критических значений коэффициента ранговой корреляции, таблица величин $-p(x_i)\log_2 p(x_i)$, необходимых для расчета энтропии распределения и количества информации. В конце книги дан предметный указатель.

В учебнике нет контрольных вопросов, решений типовых задач, контрольных заданий. Все это авторы включили в практикум, который составляет методическое сопровождение учебника.

Авторы считают своим долгом выразить благодарность рецензентам книги — коллективу кафедры общей теории статистики Московского государственного университета экономики, статистики и информатики (МЭСИ) и профессору кафедры статистики МГУ им. М. В. Ломоносова, доктору экономических наук Г. Л. Громыко за советы и замечания. Хоте-

издания

главам подготовлены И. И. Елисеевой.

[1]Вопросы статистики. — 1995. — № 12. — С. 30—32.

Глава

1

ПОНЯТИЕ О СТАТИСТИКЕ

1.1. Что такое статистика

Слово «статистика» употребляется в нескольких значениях: прежде всего как синоним слова «*данные*». Именно в этом смысле говорят: «статистика рождаемости и смертности в России» или «статистика преступлений». В этом смысле статистика входит в разделы самых различных естественных и технических наук, поскольку они связаны со сбором и обработкой массовых наблюдений, опытов и экспериментов. Соответственно можно сказать: «мне не хватает статистики» или, наоборот, «я располагаю хорошей статистикой». Мы окружены количественными данными о погоде (много раз в день получаем информацию о температуре воздуха, атмосферном давлении, направлении ветра, осадках и облачности), результатах спортивных матчей, игр, рейтингах политических деятелей и т.д.

Статистикой называется *отрасль знаний*, объединяющая принципы и методы работы с числовыми данными, характеризующими массовые явления. В этом смысле статистика включает в себя несколько самостоятельных дисциплин: общую теорию статистики как вводный курс, теорию вероятностей и математическую статистику как науки об основных категориях и математических свойствах генеральной совокупности (универсума) и их выборочных оценках.

Статистикой называют также *отрасль практической деятельности*, направленную на сбор, обработку, анализ и публикации статистических данных, отражающих явления и процессы общественной жизни. В России, как и в большинстве стран, эту работу выполняют и возглавляют специальные государственные учреждения (гл. 2).

Слово «статистика» происходит от латинского слова *status* — состояние, положение вещей. Первоначально оно употреблялось в значении «политическое состояние». Отсюда итальянское слово *stato* — государство и *statista* — знаток государства. В научный обиход слово «статистика» вошло в XVIII в. и первоначально употреблялось в значении «государствоведение».

В настоящее время *статистика может быть определена как собирание массовых данных, их обобщение, представление, анализ и интерпретация.* Это особый метод, который используется в различных сферах деятельности, в решении разнообразных задач.

Исторически развитие статистики было связано с развитием государств, с потребностями государственного управления. Хозяйственные и военные нужды уже в древний период истории человечества требовали наличия данных о населении, его составе, имущественном положении. С целью налогообложения организовывались переписи населения, проводился учет земель и т.д. Первые работы такого рода отмечены даже в священных книгах разных народов. В античном мире был организован учет родившихся (свободных граждан); молодые люди, достигшие 18 лет, вносились в списки военнообязанных, а по достижении 20 лет — в списки полноправных граждан. Составлялись земельные кадастры, в которые вносились сведения о строениях, рабах, скоте, инвентаре, доходах. Появились описания государств. Большая заслуга в этом принадлежит греческому философу Аристотелю (384—322 г. до н.э.); он составил описание 157 городов и государств своего времени.

Средневековье оставило уникальный памятник — «Книгу страшного суда» (1061 г.). Это свод материалов всеобщей переписи населения Англии и его имущества (включает данные о 240 тыс. дворов). Со временем сбор данных о массовых общественных явлениях приобрел регулярный характер; с середины XIX в. благодаря усилиям великого бельгийца — математика, астронома и статистика Адольфа Кетле (1796—1874) были выработаны правила переписей населения и регулярности их проведения. Во второй половине XVIII в. в Германии статистика была введена в университетское образование как самостоятельная учебная дисциплина. Для координации развития статистики по инициативе А. Кетле проводились международные статистические конгрессы — первый МСК состоялся в 1853 г., последний — в 1872 г.; всего было проведено 9 конгрессов. В 1885 г. был основан Международный статистический институт, существующий и сейчас. Международной статистикой занимаются международные организации — Организация Объединенных Наций (ООН), Продовольственная и сельскохозяйственная организация ООН (ФАО), Организация Объединенных Наций по вопросам образования, науки и культуры (ЮНЕСКО), Международная организация труда (МОТ), Евростат, Мировой банк, Международный валютный фонд (МВФ) и др. Международные организации и государственная статистика каждой страны занимаются сбором, представлением, интерпретацией социально-экономических данных и сравнением. Сложились методы работы, продолжающие традиции государствоведения.

Другие разделы статистики были развиты при анализе азартных игр (подсчет игровых шансов), изучении процессов воспроизводства населения. Эти достаточно сложные методы, основанные на теории вероятностей, нашли применение прежде всего в страховании и биологии, затем в других естественных науках, психологии и, наконец, с начала XX в. — в социально-экономических исследованиях, в изучении уровня жизни населения, покупательского спроса, качества продукции и т.д. Статистика нужна для расчета страховых тарифов, оценки финансовых и предпринимательских рисков; она используется в работе аудитора, при постановке управленческого учета в фирме, в контроле и анализе качества продукции, в медицине, спорте и маркетинге. Может быть, только в области искусства статистика не нашла пока широкого применения.

При изучении разных объектов в разных задачах, конечно же, используются различные методы. Тем не менее существуют некоторые общие принципы и методы статистической работы. В учебнике «Теория статистики» английских статистиков Дж. Э. Юла и М. Дж. Кендэла говорится: «Независимо от того, в какой отрасли знания получены числовые данные, они обладают определенными свойствами, для выявления которых может потребоваться особого рода научный метод обработки. Последний известен как статистический метод или, короче, статистика».

Статистические методы включают как простые методы, которые могут быть понятны любому человеку, так и сложные математические процедуры, доступные специалистам.

Различная сложность статистических методов определяет структуру статистической науки.

Статистическая наука включает:

- общую теорию статистики — изложение общих правил сбора и обработки массовых данных;
- теорию вероятностей — науку о свойствах генеральной совокупности бесконечно большого объема (так называемого универсума);
- математическую статистику, рассматривающую правила оценивания параметров и свойств генеральной совокупности по данным выборки;
- социально-экономическую статистику и статистику населения.

Поскольку лучше идти от простого к сложному, начинать изучение статистики нужно с общей теории, а потом переходить к теории вероятностей, математической статистике. В системе статистических дисциплин можно встретить «прикладную статистику», которую ведущие отечественные статистики-математики С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин определяют как «самостоятельную научную дисциплину, разрабатывающую и систематизирующую понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки (в том числе с помощью ЭВМ) статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов». Прикладная статистика по содержанию шире математической статистики, под методами которой, строго говоря, принято понимать лишь те методы статистической обработки исходных данных, разработка и использование которых апеллируют к вероятностной природе этих данных. Следовательно, прикладная статистика включает в себя часть методов математической статистики, а также математические методы описательной статистики, адаптирует порой очень сложные методы и алгоритмы применительно к типичным задачам.

Экономическая, социальная статистика, статистика населения, опираясь на общеметодологическую основу статистики, включает и специфические методы (индексный метод, балансовый метод, табличный метод, системы средних и относительных величин). Эти дисциплины могут быть детализированы по разделам экономики, социологии и демографии.

Все статистические науки тесно взаимосвязаны, частично перекрываются, но все же сохраняют самостоятельность.

В этом учебнике рассматриваются простые методы, часто используемые при изучении социальных и экономических явлений и процессов.

1.2. Статистическая закономерность. Статистические совокупности

Статистика позволяет выявить и измерить закономерности развития социально-экономических явлений и процессов, взаимосвязи между ними. Познание закономерностей возможно только в том случае, если изучаются не отдельные явления, а *совокупности* явлений — ведь закономерности общественной жизни проявляются в полной мере лишь в массе явлений. В каждом отдельном явлении *необходимое* — то, что присуще всем явлениям данного вида, проявляется в единстве со *случайным*, индивидуальным, присущим лишь этому конкретному явлению. Так, реклама какого-либо товара может не оказать влияния на рост объема продажи этого товара, однако обобщение данных о затратах на рекламу товаров и объеме их реализации показывает наличие прямой связи между этими показателями. Поэтому рекламу и называют «двигателем торговли».

Закономерности, в которых необходимость неразрывно связана в каждом отдельном явлении со случайностью и лишь во множестве явлений проявляет себя как закон, называются *статистическими*.

Кто дольше живет — мужчины или женщины? Можно привести случаи долгожительства мужчин: например, имеется свидетельство, что англичанину Фоме Карне, родившемуся в 1588 г., удалось прожить 207 лет. Абсолютно точно известно, что азербайджанец Ширали Мислимов прожил 168 лет (1805—1973). Однако это лишь частные примеры. Только при обобщении данных по всему населению выявляются закономерные соотношения. Изучая ожидаемую продолжительность жизни при рождении, видим, что при всем различии в уровне развития стран, их культуры, времени расчета

показателя общим является большая ожидаемая (и фактическая) продолжительность жизни женщин (табл. 1.1).

Таблица 1.1

Ожидаемая продолжительность жизни при рождении в 1999 г.

Страна	Мужчины	Женщины
Россия	59,9	72,4
Австрия	74,7	80,9
Великобритания	75,0	80,0
Венгрия	66,8	75,4
Германия	74,3	80,6
Финляндия	73,7	81,0
Франция	74,5	82,3
Китай	68,3	72,5

Источник. Демографический ежегодник России. 2002. Стат. сборник. — М.: Госкомстат России, 2002. — С. 106, 391—392.

Понятию статистической закономерности противостоит понятие *динамическая закономерность*, проявляющаяся в отдельном явлении. Так, площадь круга меняется с изменением его радиуса, и эта связь выражается формулой $s = 2\pi r^2$, которая справедлива для любого круга.

Свойство статистических закономерностей проявляется лишь в массе явлений при обобщении данных по достаточно большому числу единиц; оно получило название *закон больших чисел*.

Соответственно предметом статистического изучения всегда выступают *совокупности* тех или иных явлений, *включающие все множество проявлений исследуемой закономерности*. В большой совокупности индивидуальные разнообразия взаимно погашаются, и на первый план выступают закономерные свойства. Это мы уже видели на примере сравнения ожидаемой продолжительности жизни мужчин и женщин. То же наблюдается и для экономических явлений. Так, цены на отдельные товары могут понижаться, на другие — повышаться, но совокупное изменение цен на все потребительские товары и услуги свидетельствует о неуклонном росте цен. Статистические совокупности часто называют *массовыми явлениями*.

Статистические закономерности обладают свойством *устойчивости*, т.е. стабильности и повторяемости при повторных наблюдениях.

В течение более или менее длительного промежутка времени характеристики остаются примерно постоянными. Так, доля мальчиков и девочек среди новорожденных колеблется слабо: 105—106 мальчиков на 100 новорожденных девочек; доля лиц разных возрастов среди вступающих в брак и т.д. обнаруживает от года к году не очень значительные колебания. Эти факты представляют громадный интерес. Устойчивость определяет возможность существования и развития общества, на этом свойстве базируются прогнозы, скажем, прогноз пропорций между отраслями и секторами экономики и т.д.

Каждое единичное явление рассматривается статистикой как особый, частный случай изучаемой закономерности. Статистика дает количественную характеристику исследуемой закономерности, а это возможно лишь при обобщении всего множества ее проявлений, взятых в целом, т.е. на основе совокупности явлений. Количественная характеристика каждого отдельного явления отражает его сущность. Но эта частная характеристика ограничена в своем значении для познания закономерности, поскольку сложилась в конкретных условиях и в силу этого соединяет в себе как типичные черты, присущие всем явлениям данного вида, так и случайные, присущие именно этой конкретной единице. Например, в Санкт-Петербурге имеются крупные семьи, включающие родителей и 5—6 детей; бездетные семьи, состоящие лишь из мужа и жены, но и те, и другие не являются типичными для этого города. Только обобщив данные по всем петербургским семьям, можно говорить о том, какой размер семьи закономерен для северной столицы на современном этапе (3,1 человека), а также о том, что типичными являются так называемые простые семьи, состоящие из родителей и одного-двух детей, без прочих родственников.

Поскольку статистика призвана выявлять закономерное, она, опираясь на данные о каждом отдельном проявлении изучаемой закономерности, обобщает их и таким образом получает количественное выражение этой закономерности.

Статистика не связана с каким-либо конкретным измерителем. Она использует как стоимостные (денежные), так и на-

туральные показатели. Для анализа динамики стоимостные показатели выражаются не только в текущих ценах, но и в так называемых неизменных ценах, т.е. ценах, установленных за определенный период или на определенную дату, применяемых в течение ряда лет для оценки продукции в отдельных отраслях материального производства.

Стоимостное выражение позволяет агрегировать данные о производстве или продаже разнородных товаров и услуг, например, рассчитывать валовую продукцию предприятия, объединения, отрасли. Кроме денег в качестве универсального соизмерителя могут использоваться затраты труда на производство товаров и услуг. Затраты труда выражаются в человеко-часах, человеко-днях. При использовании этого соизмерителя возникают проблемы из-за различий в квалификации, навыках и умении работников. При обобщении натуральных показателей могут возникнуть трудности из-за несопоставимости данных. Преодолеть их позволяют *условно-натуральные измерители*. Например, рыбные консервы выпускаются в больших и маленьких банках, высоких и низких, причем в разные годы соотношение между ними меняется. Для того чтобы подсчитать, сколько произведено консервов, сравнить эту цифру с прошлым периодом, используют так называемые условные банки. Чтобы обобщить мощность двигателей по совокупности предприятий, их выражают в лошадиных силах, а затем суммируют. Топливо разной теплотворной способности пересчитывают в условное топливо; скот (коров, быков, коз, овец и т.д.) пересчитывают в условные головы крупного рогатого скота и т.д.

Несмотря на материальные различия изучаемых статистикой совокупностей, все они имеют общие черты.

Статистическая совокупность состоит из *единиц совокупности*. Каждая единица совокупности представляет собой частный случай проявления изучаемой закономерности.

Объединение единиц в совокупность объективно обосновано, это не произвол исследователя. В самом деле, не вызывает сомнения объективность существования таких совокупностей, как машиностроительные предприятия, продовольственные магазины, население страны и другие, которые изучает социально-экономическая статистика. Как бы далеко друг от друга ни находились единицы каждой из перечисленных совокупностей, они взаимосвязаны. В их существовании,

взаимосвязях, развитии формируются соответствующие закономерности и тенденции развития машиностроения, торговли продовольственными товарами, воспроизводства населения и его структуры и т.д.

Социально-экономические явления отличаются особенно сложной природой. В каждом отдельном явлении одновременно реализуются различные процессы. Например, работник может рассматриваться как член определенной социально-профессиональной группы, представитель коллектива работников предприятия, на котором он трудится, часть населения того города, поселка, где он живет, и т.д.

Важнейшая особенность «включенности» единиц в разные процессы состоит в том, что как члены той или иной совокупности они выступают лишь *в одной связи*, в аспекте одного процесса. Так, если изучаются численность и состав определенной социально-профессиональной группы, работник рассматривается как единица совокупности, образуемой промышленно-производственным персоналом предприятия, и т.д. Если же изучается воспроизводство населения, то тот же человек рассматривается как часть населения и нас интересуют прежде всего его семейное положение (одиночка или живет в семье), возраст, наличие детей и пр. Таким образом, решение вопроса о единице и границах изучаемой совокупности зависит от цели исследования. Если, например, изучается население как основа формирования трудовых ресурсов, то единицей совокупности будет человек, тогда как при изучении потребления населением единицей является домохозяйство как потребительская ячейка.

Многие социально-экономические проблемы носят комплексный характер. Их исследование требует совместного рассмотрения разных совокупностей. Так, изучение процесса воспроизводства населения предполагает анализ всех основных процессов, в которые вовлечен, с одной стороны, человек как единица совокупности, с другой — семья. Ведь основные характеристики демографического и социального воспроизводства населения зависят не только от структуры населения, но и от состава семей: наличия брачной пары, возраста супругов, наличия детей, прочих родственников.

При одной и той же цели исследования особенности решения вопроса о единице и соответственно об изучаемой со-

вокупности зависят еще и от уровня исследования. Можно изучать, например, производительность труда на уровне отрасли, отдельного предприятия, цеха, бригады, наконец, отдельного рабочего. В каждом случае единица совокупности будет особой: предприятие данной отрасли; рабочий данного предприятия, цеха, бригады; отработанный человеко-день (или человеко-час) — при изучении выработки отдельного рабочего. Уровень исследования определяет круг выдвигаемых задач, и, наоборот, задачи исследования определяют уровень его организации. В том, как указана единица совокупности, проявляется непосредственная связанность этих вопросов. При исследовании на любом уровне в качестве единиц выступает то явление, в котором реализуется изучаемая закономерность, наблюдая за которым, можно проследить ее действие (в той мере, в какой это возможно в единичном явлении).

Такой подход приводит к еще одному определению единицы: *единица совокупности — это предел дробления объекта исследования, при котором сохраняются все свойства изучаемого процесса.*

Иногда бывает довольно трудно логически обосновать единицу совокупности по той причине, что отсутствуют «естественные» пределы дробления. Например, при изучении влияния удобрений на урожайность определенной культуры в качестве единицы могут выступать отдельный массив посевов (поле или делянка), бригада, сельскохозяйственное предприятие, район и даже республика. Такая многозначность решений возникает, к примеру, если проводить исследование исключительно с точки зрения природных условий. Если же проводить изучение с точки зрения экономических и организационных факторов, то предельным уровнем дробления является сельскохозяйственное предприятие (ферма, товарищество и т.д.).

Итак, предметом статистического изучения выступают *совокупности — множество однокачественных, варьирующих явлений.* В это определение входят три основные черты совокупности любых явлений: во-первых, это множество явлений; во-вторых, это множество явлений, объединенных общим качеством, представляющих собой проявления одной и той же закономерности; в-третьих, это множество варьирующих явлений, отличающихся по своим характеристикам.

Именно последнее свойство вызывает необходимость изучения всего множества явлений одного вида. Если бы единицы совокупности были полностью тождественны друг другу, то не было бы потребности обращаться к множеству единиц: достаточно изучить лишь одну единицу, чтобы знать все о всех явлениях этого вида.

Вариация — основа существования мира и источник его развития. Если бы люди не делились на мужчин и женщин, человечество прекратило бы существование; если бы не было различных мнений — истина была бы недостижимой, а жизнь без вариаций — невыносимо скучной!

Подводя итог сказанному, отметим, что предметом статистического изучения могут выступать данные:

пространственные (число единиц совокупности велико, данные относятся к одному времени: $N \to \infty$, $t = 1$);

панельные: $N \to \infty$, $1 < t \leqslant 10$ (близкие к пространственным данным);

временные ряды: $N = 1$, $t \to \infty$.

1.3. Признаки и их классификация

Единицы совокупности обладают определенными свойствами, качествами. Эти свойства принято называть *признаками.* Например, признаки человека: возраст, образование, занятие, рост, вес, семейное положение и т.д.; признаки предприятия: форма собственности, специализация (отрасль), численность работников, величина уставного фонда, экономическая эффективность его деятельности и т.д.

Статистика изучает явления через их признаки: чем более однородна совокупность, тем больше общих признаков имеют ее единицы и тем меньше варьируются их значения.

Признаки различаются способами их измерения и другими особенностями, влияющими на приемы статистического изучения. Это дает основание для классификации признаков (табл. 1.2).

Описательные признаки выражаются словесно: национальность человека, разновидность почв, материал стен здания. Описательные признаки подразделяются на *номинальные* и *порядковые.* Эти термины взяты из теории измерений. Отличия между ними в том, что номинальные — это описатель-

Таблица 1.2

Классификация признаков в статистике

Основы классификации				
по характеру выражения	по способу измерения	по отношению к характеризуемому объекту	по характеру вариации	по отношению ко времени
Описательные	Первичные, или учитываемые	Прямые (непосредственные)	Альтернативные	Моментные
Количественные	Вторичные, или расчетные	Косвенные	Дискретные	Интервальные
			Непрерывные	

ные признаки, по которым нельзя ранжировать (упорядочивать) данные, а порядковые — это признаки, по которым можно ранжировать данные. Например, пользуясь оценками экспертов, ранжируют фигуристов по технике и артистичности исполнения программы или работников — по мастерству, студентов — по успеваемости и т.д.

Количественные признаки выражаются числами. Они играют главенствующую роль в статистике. Таковы возраст человека, площадь пашни, заработная плата рабочих, население города, доход кооператива и т.д.

Первичные признаки характеризуют единицу совокупности в целом. Это абсолютные величины. Они могут быть измерены, сосчитаны, взвешены и существуют сами по себе независимо от их статистического изучения. Например, площадь пашни, мощность двигателей на предприятии, численность населения города, число автомобилей, произведенных в стране.

Вторичные, или *расчетные*, признаки не измеряются непосредственно, а рассчитываются. Они являются продуктами человеческого сознания, результатом познания изучаемого объекта. Например, себестоимость единицы продукции, производительность труда, рентабельность, урожайность и т.п. Вторичные признаки представляют собой соотношения первичных признаков: деление объема выпущенной продукции на численность работников дает показатель производительности труда; деление суммы затрат на произведенную продукцию на число единиц данной продукции дает себестоимость и т.д. Несмотря на расчетный характер, вторичные признаки тоже имеют объективный характер. Процесс познания есть отражение объективных свойств явлений и процессов, и расчеты, статистические методы познания являются таким же необходимым средством отражения объективных свойств совокупности, как измерение, взвешивание. Вторичный — не означает второстепенный. Термин определяет только путь познания: сначала надо измерить значения первичных признаков, а уже потом, во вторую очередь, на основе первичных признаков рассчитать значения вторичных.

Прямые (*непосредственные*) признаки — это свойства, непосредственно присущие тому объекту, который ими характеризуется. Таковы возраст человека, поголовье коров на ферме, объем продукции завода, численность его рабочих.

Косвенные признаки являются свойствами, присущими не самому объекту, а другим совокупностям, относящимся к объекту, входящим в него. Например, продуктивность коров как косвенный признак фермы. Хотя продуктивность не фермы, а коров — это их прямой признак, но ведь продуктивность характеризует и ферму, которой принадлежат эти коровы (или даже целую область). Такова и оплата труда рабочих по отношению к заводу. Это косвенный признак завода, но очень важный для того, кто собирается поступать на работу и выбирает предприятие.

Практически деление признаков на прямые и косвенные совпадает с их делением на первичные и вторичные.

Признаки различаются в статистике и по характеру их вариации, т.е. по различиям их значений у разных единиц совокупности. Выделяются *альтернативные* признаки, которые могут принимать только два значения. Таковыми являются признаки обладания или необладания чем-то. Например, все садовые участки по признаку наличия посадок вишни можно разделить на имеющие посадки вишни и не имеющие их. Альтернативным признаком являются пол человека, место проживания (город, село), ходовая система трактора (гусеничный или колесный).

К *дискретным* относятся количественные признаки, которые могут принимать только целочисленные значения, без

промежуточных значений между ними. Это число членов семьи, количество этажей здания, комнат в квартире.

Непрерывные, точнее непрерывно варьирующиеся, признаки способны принимать любые значения, конечно, в определенных границах. К непрерывным относятся расчетные вторичные признаки. Ведь их значения — результат деления, а оно может приводить к любым числам — целым, дробным, иррациональным. На практике нередко значения непрерывных признаков округляют с конечной степенью точности, так что они становятся квазидискретными. С другой стороны, дискретные по существу признаки, например число работников предприятия на 1 января, поголовье коров на ту же дату, имеют такое громадное число возможных значений, что на практике статистика вынуждена обращаться с ними, как с квазинепрерывными. Об этом будет сказано в главах 5 и 6 при обсуждении метода группировок и расчета средних величин.

Моментные признаки характеризуют изучаемый объект в какой-то момент времени, установленный планом статистического исследования. Они существуют на любой момент и характеризуют наличие чего-либо: численность населения, стоимость фондов, количество скота, размеры жилой площади.

К *интервальным* относятся признаки, характеризующие результаты процессов. Поэтому их значения могут возникать только за интервал времени: год, месяц, сутки, но не на момент времени. Таковы число родившихся, умерших, объем промышленной продукции, надой молока, сумма полученной прибыли. Моментные признаки — характеристики *состояния*, а интервальные — характеристики *процесса*. Различие между моментными и интервальными признаками существенно при изучении динамики. Единицы измерения моментных признаков относятся только к характеризуемым ими свойствам объектов, а единицы измерения интервальных признаков содержат еще и указание того отрезка времени, за который определено значение признака. Так, стоимость основных производственных фондов предприятия на 1 января выражается в миллионах рублей, а объем продукции за январь — в тысячах или миллионах рублей за месяц.

1.4. Определение предмета статистики — основа статистической методологии

Как уже отмечалось, предметом статистического изучения всегда выступает совокупность явлений. Как правило, она включает в себя несколько частных совокупностей, представляющих особые типы явлений, иначе говоря, особые модификации изучаемой закономерности. Единицы разных частных совокупностей в рамках общего качества отличаются кругом признаков и их значений.

В большинстве случаев правильным будем представление частной совокупности (однородной группы), состоящей из ядра и окружающих его явлений — слоя. Ядро — концентрированное выражение всех специфических свойств типа (группы), определяющих качественное отличие данного типа от всех иных. Кроме единиц, составляющих ядро, тип включает явления переходного качества («слой»), принадлежность которых к данному типу может быть установлена с определенной вероятностью. Подобные явления образуют, так сказать, «полосу размыва» между типами.

Среди студентов можно встретить тип «идеальный студент»: прекрасно учится, много читает, хороший товарищ. Есть студенты не такие разносторонние, для которых важны только специальные знания; есть и другие типы. «Качество» одних студентов, их принадлежность к тому или иному типу можно определить практически безошибочно, тогда как других бывает трудно отнести к какому-то типу. Они-то и представляют собой явления переходного качества.

Соотношение между ядром и его окружением в разных типах будет, конечно, различным: это зависит от устойчивости типа, длительности его существования, взаимодействия с другими типами той же совокупности, с другими совокупностями. Однако ядро должно составлять большинство единиц того или иного типа, так как именно ядро определяет «лицо» типа, его характерные свойства.

Социально-экономическая статистика изучает совокупности однокачественных явлений в конкретных условиях места и времени. Таким образом, статистика располагает всегда ограниченным числом данных. Каждое явление возникает как результат множества факторов. В естественных науках

можно проследить интересующие взаимосвязи с помощью специально проведенных лабораторных экспериментов, которые называют активными экспериментами, так как исследователь практически полностью контролирует ход эксперимента и может выделить в более или менее чистом виде влияние каждого из выбранных факторов, элиминируя влияние остальных. Иная ситуация в социально-экономических исследованиях. «При анализе экономических форм нельзя пользоваться ни микроскопом, ни химическими реактивами. То и другое должна заменить сила абстракции», — писал К. Маркс (Маркс К., Энгельс Ф. Соч. — 2-е изд. — Т. 23. — С. 4). Применяя различные методы анализа, мы проводим «пассивный» эксперимент, причем ни один метод не позволяет определить «чистый» вклад каждого из факторов по отдельности в совокупный результат.

В центре социально-экономических явлений и процессов находится человек со своими субъективными установками, активным воздействием на окружающий мир; это делает достоверность данных важнейшей проблемой статистики.

Обобщая сказанное, можно указать следующие *особенности социально-экономических явлений*: 1) сложность их материальной природы, многообразие количественных и качественных определений; 2) ограниченность численности; 3) динамичность; 4) многообразие видов и форм, в которых проявляются единые по своей сущности процессы, отсюда — разделение на частные совокупности, на группы особого качества; 5) взаимосвязанность явлений и признаков; невозможность элиминирования действия факторов и раздельной оценки их действия.

Специфика предмета статистики обусловливает специфику *статистического метода*. Он включает *сбор данных* (статистическое наблюдение), их *обобщение, представление, анализ и интерпретацию*.

Статистические данные могут быть взяты из публикаций, а можно собрать новую информацию по каждой единице совокупности (фирме, человеку, виду продукции, товару). Получение исходных данных является одной из наиболее трудных и важных задач, которые встают перед статистикой. Главное — использовать те данные, которым можно доверять.

Обобщение данных наблюдения включает группировку — разграничение общей совокупности на группы однородных единиц и сводку — обобщение значений признаков в сводные статистические показатели для характеристики каждой частной совокупности, группы и совокупности в целом (гл. 3, 5, 6). Часто данные можно получить лишь выборочным методом, а затем по выборке составить суждение о генеральной совокупности, из которой формировалась выборка. Нередко приходится идти путем испытания статистических гипотез, выдвигая предположения о свойствах генеральной совокупности и проверяя их с помощью статистико-математических критериев (гл. 7 и 8).

Для того чтобы пользоваться результатами обобщения или непосредственно исходной информацией, данные должны быть представлены в подходящей форме, компактно и наглядно. С этой целью строятся таблицы и графики (гл. 4).

Процесс анализа охватывает все стадии статистического исследования. Каждый следующий этап статистической работы зависит от предыдущего. Этап обобщения данных оказывает влияние на статистическое наблюдение — ведь именно тем, что мы хотим получить в результате исследования, определяются границы объекта наблюдения, программа наблюдения (какие признаки мы будем регистрировать у единиц совокупности).

Выделение типов в результате классификации или группировки данных обеспечивает их однородность. Тем самым создается основа для расчета сводных показателей, анализа вариации и связей. Однородность обобщаемых данных определяет устойчивость всех статистических показателей. Например, очевидно, что устойчивость значения среднего надоя молока будет разной в том случае, если показатель рассчитан в целом по России или по отдельным территориям, скажем, федеральным округам с достаточно однородными природно-климатическими условиями.

При изучении связей статистика помогает установить круг важнейших факторов, измерить хотя бы и условно силу их влияния (гл. 9, 11). В решении этой задачи всегда существует опасность установления ложных связей — принять за причину просто сопутствующие явления. Например, считать черного кота или разбитое зеркало предвестием неудач.

Важным направлением анализа является изучение динамики. Чтобы предсказать развитие в будущем (сколько автомобилей будет произведено и продано на внутреннем рынке, какова будет численность населения в 2005 г. и т.д.), нужно знать фактическую динамику в прошлом: как изменялись показатели, имелась ли тенденция в их изменении, каков характер колеблемости данных.

Каждый шаг исследования завершается интерпретацией полученных результатов: какое заключение можно сделать исходя из проведенного анализа, что говорят нам цифры — подтверждают ли они исходные предположения или открывают что-то новое? Интерпретация данных ограничена исходным материалом. Если заключения основаны на данных выборки, то она должна быть репрезентативной, чтобы выводы были отнесены к совокупности в целом (гл. 7). Статистика позволяет выяснить все то полезное, что содержится в исходных данных, и определить, что и как можно использовать в принятии решений.

РЕЗЮМЕ

Термин «статистика» может означать массовые данные, отрасль знаний, область профессионального занятия.

Статистика выделилась как самостоятельная наука во второй половине XVIII в.

Статистика — наука о методах сбора, представления, обработки и анализа данных. Статистические методы адаптируются к изучаемым явлениям.

Статистическая наука включает общую теорию статистики (дескриптивную статистику), теорию вероятностей, математическую статистику. Возможны выделения и других разделов этой области знания.

Предмет статистики — статистическая совокупность, т.е. множество однокачественных варьирующих явлений. Могут изучаться пространственные, панельные, временные данные.

В статистической совокупности реализуется статистическая закономерность, которая проявляется при обобщении множества явлений. Это свойство статистической закономерности получило название закона больших чисел. Статистическая закономерность обладает устойчивостью, повторяемостью.

Категории предмета статистического изучения:

- частная совокупность, или особый тип явлений;
- структура типа: ядро и слой (промежуточные явления);
- единица совокупности — частный случай изучаемой закономерности;
- признак — свойство единицы совокупности;
- показатель — характеристика группы явлений.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Айвазян С., Мхитарян В.* К концепции статистического образования экономистов // Вопросы статистики. — 1995. — № 12. — С. 34—38.
2. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. — М.: Финансы и статистика, 1983.
3. *Елисеева И. И.* Моя профессия — статистик. — М.: Финансы и статистика, 1992.
4. *Плошко Б. Г.* Группировка и системы статистических показателей. — М.: Статистика, 1978.

Глава

2

ОРГАНИЗАЦИЯ СТАТИСТИКИ. СТАТИСТИЧЕСКОЕ НАБЛЮДЕНИЕ

2.1. Организация государственной статистики в Российской Федерации

Основная продукция статистики — это обобщающие статистические показатели, такие, как численность населения в стране (или регионе), процент женщин, процент лиц с высшим образованием, валовой внутренний продукт, произведенный за год (или квартал), и т.д. Для того чтобы получить такую обобщенную информацию, нужно организовать сбор данных (статистическое наблюдение), их обобщение и представление в компактной и наглядной форме. Статистические показатели агрегируют данные массового наблюдения по совокупности единиц. Они используются правительством страны, местными органами власти; лежат в основе межправительственных соглашений. Следовательно, должна существовать *официальная государственная статистика* — система организаций, отвечающих за сбор социальных и экономических данных и их обобщение по определенной территории, за их точность и достоверность.

Система государственной статистики в России сложилась во второй половине XIX в. Главным органом государственной статистики в настоящее время является Государственный комитет Российской Федерации по статистике (Госкомстат России). В каждом субъекте РФ имеется региональный комитет по статистике. В целом можно сказать, что структура органов

государственной статистики соответствует административному делению страны. Низовым звеном являются районные или городские отделы государственной статистики, которые имеются в административных районах краев, областей, а также в административных районах крупных городов, таких, как Москва, Санкт-Петербург.

Статистические данные о численности и составе населения, о структуре экономики и результатах ее работы за период, состоянии рынка труда, уровне цен и их изменениях, заработной плате и совокупных доходах населения и т.д. публикуются в специальных изданиях — статистических сборниках. Сборники бывают двух видов: универсальные и специальные, освещающие детально состояние какой-либо сферы. Примером универсального статистического сборника могут служить краткий статистический ежегодник «Россия в цифрах в ... году» или полный «Статистический ежегодник Российской Федерации за ... год», к этому типу изданий можно отнести и двухтомник «Регионы России». Универсальные сборники издаются ежегодно Госкомстатом России.

К специализированным статистическим сборникам относятся: «Социальное положение и уровень жизни населения России» (ежегодник), «Демографический ежегодник России», «Промышленность России», «Цены в России», «Национальные счета России в ... году» и др.

Местные статистические органы издают региональные статистические сборники. Например, Петербургкомстат издает ежегодник «Санкт-Петербург в ... году» и «Ленинградская область в ... году». Комитет по статистике Москвы — статистический сборник «Москва в цифрах, ... год» (краткий), «Москва в ... году». В регионах издаются и специализированные статистические сборники.

Структура Госкомстата России и региональных комитетов по статистике соответствует основным направлениям статистических работ.

Госкомстат России включает управления:

- статистического планирования и организации статистического наблюдения;
- национальных счетов;
- статистики предприятий и структурных обследований;
- статистики труда;

- переписи населения и демографической статистики;
- статистики зарубежных стран и международного сотрудничества;
- статистики уровня жизни и обследований населения;
- статистики услуг, транспорта и связи;
- статистики основных фондов и строительства;
- статистики внутренней и внешней торговли;
- статистики окружающей среды и сельского хозяйства;
- статистики цен и финансов;
- сводной информации.

Подобные подразделения (отделы) выделяются в региональных комитетах по статистике.

Государственный комитет по статистике РФ входит в структуру федеральных органов исполнительной власти. Региональные комитеты статистики входят в структуру местных органов власти. Оперативность и качество статистических работ зависят от развития технологии сбора, передачи, обработки и хранения информации. Все областные, краевые и республиканские комитеты по статистике имеют вычислительные центры. Мощный вычислительный центр имеет Госкомстат России. Все большее значение приобретают локальные вычислительные сети, связывающие банки данных статистических служб, других держателей региональной и федеральной информации.

Большую роль в методологической работе играет Научно-исследовательский институт статистики Госкомстата России. В этой работе принимает участие и Научно-методологический совет Госкомстата России, который объединяет ведущих работников государственной статистики и представителей экономической и статистической науки.

Основные функции всех статистических органов состоят в сборе, обработке, анализе и представлении данных в удобном для пользователя виде. Статистические службы должны оперативно предоставлять информацию органам управления, осуществлять обмен информацией с Центральным банком Российской Федерации (Банком России) и его конторами на местах, Министерством финансов РФ (Минфином) и его местными органами, Министерством экономического развития, торговли и туризма, Министерством по труду и социальной защите населения РФ и т.д.

Госкомстат России является *методологическим и организационным центром* работы всех служб государственной статистики. Здесь разрабатываются федеральный план статистических работ на год и перспективу, методология расчета статистических показателей, сбора и разработки статистических данных, подводятся итоги работы государственной статистики за год. Методологическая работа Госкомстата России направлена на внедрение интегрированной системы учета и статистики, соответствующей международным стандартам, прежде всего на разработку системы национальных счетов РФ, позволяющей исследовать формирование основных пропорций экономики и рассчитывать важнейшие макроэкономические показатели, используемые в международной практике, а также на измерение инфляции и уровня жизни. Эта работа ведется при участии международных статистических организаций и национальных статистических служб развитых стран.

Широко распространились международные связи между национальными статистическими службами и на региональном уровне.

Так, Комитет по статистике Республики Карелия имеет постоянные связи и совместные проекты со статистическими органами Финляндии по направлениям «Обследования домашних хозяйств», «Регистр населения». Статистики скандинавских стран и северных регионов России издали статистический сборник «Женщины и мужчины в 1997 г.». Результатом шведско-российского проекта «Совершенствование гендерной статистики в России» стал сборник «Женщины и мужчины в Республике Карелия».

С финскими и шведскими коллегами взаимодействуют статистики Петербургкомстата, осуществляя совместные издания статистических сборников (например, «Санкт-Петербург — Хельсинки—Турку в цифрах» (1997 г.)). Результатом сотрудничества Госкомстата России с Федеральным статистическим управлением ФРГ явилось издание краткого статистического сборника «Россия и Германия в 2001 г.».

Официальная статистика в России является *централизованной:* руководство ею составляет функцию самостоятельного государственного учреждения — Госкомстата России. Но, конечно же, статистическую работу ведут все министерства и ведомства, т.е. существует *ведомственная статистика*. Спе-

циальные базы статистических данных имеются у Банка России и Минфина, Минобразования и Министерства труда и социальной защиты и т.д. Центробанк издает свои статистические публикации, например «Статистический бюллетень Центрального банка» и др.

Пользователями статистической информации выступают прежде всего органы власти: Правительство РФ, административные органы федеральных округов (Северо-Западный, Северо-Кавказский, Центральный, Поволжский, Уральский, Восточно-Сибирский, Приморский), субъектов РФ, районов. В последние годы круг пользователей статистической информацией в России непрерывно расширяется. Статистические данные (о населении, предприятиях) требуются муниципалитетам. Все чаще к статистике прибегают представители бизнеса, прежде всего маркетинговые службы для оценки мощности и структуры рынка того или иного товара, услуги. Ведь даже для того, чтобы решить, сколько хлебобулочных изделий необходимо произвести для удовлетворения спроса и чтобы не было избытка, нужно знать не просто численность жителей на данной территории, но и их возрастной состав (младенцы хлеб не едят, дети едят его меньше, чем взрослые), половой состав (мужчины потребляют хлеба больше, чем женщины), национальный состав (представители кавказских народов любят лаваш, украинцы предпочитают пышный белый хлеб, русские любят ржаной (черный) хлеб). Статистические данные нужны и коммерческим банкам, чтобы представлять возможности расширения клиентской базы, классифицировать заемщиков и т.д.

Статистику используют студенты и школьники для написания индивидуальных проектов, курсовых, дипломных работ; она необходима аспирантам при выполнении диссертационных исследований; без статистики не могут обойтись ученые. Любой гражданин может заинтересоваться статистикой. Ее широко применяют средства массовой информации (СМИ). В силу своей природы СМИ могут воздействовать на людей с помощью статистики, приводя данные о смертности, рождаемости или пугая нашествием мигрантов. Здесь очень важны культура журналистов, степень их хотя бы общего понимания тех показателей, которыми они манипулируют, чтобы не было путаницы в употреблении терминов «валовой внутренний продукт» (ВВП) и «валовой региональный продукт» (ВРП), «налог на добавленную стоимость» (НДС) и «система национальных счетов» (СНС).

Статистические организации должны исходить из потребностей пользователя и стремиться к тому, чтобы каждый нашел полезную для себя информацию. Для пользователя важно, чтобы статистический сборник содержал год от года один и тот же набор показателей, чтобы названия таблиц и их расположение были идентичными (принцип консерватизма). Тогда легко проводить сравнения в динамике. Конечно, изменения в окружающем мире также должны отражаться в сборниках, должны появляться новые показатели. Например, Петербургкомстат с 2002 г. публикует данные о количестве пользователей Интернетом в регионе, а с 2001 г. — о количестве пользователей сотовой и пейджинговой связью.

Статистические службы в регионах выполняют тот объем работ, который соответствует плану статистических работ, разработанному Госкомстатом России. Эти работы финансируются из федерального бюджета. К таким работам относится, например, Всероссийская перепись населения 2002 года. Но кроме того, возникает потребность в дополнительных разработках, для отражения каких-то специфических особенностей региона, освещения специальных проблем. Эти работы должны выполняться за счет местного бюджета. Показатели, рассчитываемые для региона по программе Госкомстата, в совокупности с показателями, рассчитанными за счет средств местного бюджета, образуют *региональную статистику*. Администрация любого региона интересуется, например, ходом жилищно-коммунальной реформы, и нужно знать доходы населения, какую долю в расходах домохозяйств составляет оплата жилищно-коммунальных услуг, сколько владельцев жилищ имеют льготы, категории льготников, сколько граждан должны оплачивать жилищно-коммунальные услуги в полном объеме, какова задолженность по оплате жилищно-коммунальных услуг и т.д. И так по каждой региональной программе.

Источником статистических сведений могут быть не только официальные органы, но и исследовательские группы («альтернативная статистика»). Чаще всего это социологические институты и службы, организующие опросы населения, результаты которых доводятся как до граждан, так и до исполнительных и законодательных органов.

2.2. Важнейшие международные организации и их статистические службы

В развитых странах национальные службы, занимающиеся статистикой, сложились к первой половине XIX в. С тех пор стала четко осознаваться потребность в сопоставимости данных разных стран, выработке единых рекомендаций.

Первый шаг в решении этих задач был сделан бельгийским статистиком Адольфом Кетле (1796—1874). По его инициативе возникло международное сотрудничество в области статистики — Международный статистический конгресс (МСК). Состоялось девять сессий МСК (восемь было проведено А. Кетле): в 1853 г. (Брюссель), 1855 г. (Париж), 1857 г. (Вена), 1860 г. (Лондон), 1863 г. (Берлин), 1867 г. (Флоренция), 1869 г. (Гаага), 1872 г. (Санкт-Петербург), 1876 г. (Будапешт). В них принимали участие работники административной статистики, деятели статистической науки и представители общественности разных стран. Обсуждались вопросы организации административной статистики, ее централизации, проведения переписей населения, классификации производств и т.д. Международные статистические конгрессы способствовали широкому обмену опытом и заметно повысили уровень статистических работ.

Следующая стадия развития международной статистики связана с деятельностью Международного статистического института (МСИ). В 1985 г. отмечалось 100-летие МСИ, который продолжает успешно работать и по сей день. Регулярно каждые два года проходят сессии МСИ (первая состоялась в 1887 г.). В 1913 г. для руководства работой МСИ было создано Постоянное бюро МСИ, которое с 1923 г. публикует ежемесячные бюллетени. С 1933 г. издается журнал МСИ, с 1975 г. публикуются ежегодные отчеты.

Международный статистический институт является международной научной организацией. В его задачи входят развитие официальной статистики, унификация методов сбора и способов разработки статистических данных в разных странах, приведение публикаций к единообразию для возможности международных сравнений, развитие статистической науки и образования, компьютерной поддержки. С момента организации в работе МСИ принимают участие не только ведущие ученые-статистики, но и практики.

Современный этап международной статистики связан с деятельностью Организации Объединенных Наций (ООН), Международного валютного фонда (МВФ), Всемирного банка (ВБ).

Организация Объединенных Наций была создана в 1945 г. для поддержания мира, развития сотрудничества и дружеских отношений между странами. В ее состав входят свыше 160 государств. Штаб-квартира ООН находится в Нью-Йорке, а европейские отделения — в Женеве и Вене. Секретариаты региональных комиссий находятся в Аддис-Абебе, Бангкоке, Багдаде и Сантьяго.

В соответствии с уставом ООН ее главными органами являются: Генеральная Ассамблея, Совет Безопасности, Экономический и социальный совет, Совет по опеке, Международный суд и Секретариат.

Секретариат ООН обеспечивает текущую работу главных и вспомогательных органов ООН, отвечает за выполнение их решений. Возглавляет Секретариат Генеральный секретарь ООН, избираемый на пятилетний срок.

Экономический и социальный совет (ЭКОСОС) ежегодно проводит две сессии продолжительностью около месяца по различным направлениям народного хозяйства, одну — в Нью-Йорке, другую — в Женеве. В период между сессиями осуществляет свою деятельность через постоянные комитеты, комиссии. В ведении Совета находятся пять региональных комиссий, которые изучают положение дел в своих регионах и помогают правительствам выработать политику для решения стоящих перед ними проблем. На постоянно действующей основе ЭКОСОС имеет:

- функциональные комиссии (статистическую комиссию, комиссии по народонаселению, по правам человека, по положению женщин и т.д.);
- комитеты (комитет по программе и координации, комитеты по природным ресурсам, по неправительственным организациям, по транснациональным корпорациям и т.д.);
- органы экспертов (межправительственные рабочие группы экспертов по международным стандартам учета и ста-

тистики, по международному сотрудничеству в области налогообложения, по перевозке опасных грузов и т.д.);
- региональные комиссии.

Специализированные учреждения ООН — совокупность межправительственных учреждений и организаций, связанных с ООН специальными соглашениями. Они взаимодействуют с ООН и друг с другом через Экономический и социальный совет и Административный совет по координации (АКК), возглавляемый Генеральным секретарем ООН. Таких учреждений 16.

1. Организация Объединенных Наций по вопросам образования, науки и культуры (ЮНЕСКО).
2. Всемирная организация здравоохранения (ВОЗ).
3. Международный банк реконструкции и развития (МБРР).
4. Международный валютный фонд (МВФ).
5. Международная организация труда (МОТ).
6. Продовольственная и сельскохозяйственная организация ООН (ФАО).
7. Международная ассоциация развития (МАР).
8. Международная финансовая корпорация (МФК).
9. Международная организация гражданской авиации (ИКАО).
10. Всемирный почтовый союз (ВПС).
11. Международный союз электросвязи (МСЭ).
12. Всемирная метеорологическая организация (ВМО).
13. Международная морская организация (ИМО).
14. Всемирная организация интеллектуальной собственности (ВОИС).
15. Международный фонд сельскохозяйственного развития (МФСР).
16. Организация Объединенных Наций по промышленному развитию (ЮНИДО).

Генеральное соглашение по тарифам и торговле (ГАТТ), Международное агентство по атомной энергии (МАГАТЭ) и Международный торговый центр ЮНКТАД-ГАТТ (МТЦ) не являются специализированными учреждениями, но входят в общую систему ООН.

В главных органах Организации Объединенных Наций работают два крупных статистических учреждения: Статистическая комиссия в структуре ЭКОСОС и Статистическое бюро в составе Секретариата ООН. Дополнительно созданы статистические службы во всех региональных комиссиях и специализированных учреждениях, работающих в тесной связи между собой и другими международными статистическими службами (рис. 2.1).

Статистическая комиссия ООН руководит методологической работой, обобщает и анализирует статистический опыт отдельных стран, занимается непосредственно методикой и методической работой с целью повышения сопоставимости данных по различным территориям, периодам и организациям, разрабатывает статистические стандарты, оказывает консультативную помощь по вопросам статистики, координирует статистическую работу исполнительных и специализированных учреждений ООН, оказывает помощь в сборе, обработке, хранении и передаче информации.

Итоговой формой работы Статистической комиссии ООН считается сессия, которая проводится один раз в два года.

В основу деятельности Статистического бюро секретариата, который является исполнительным статистическим органом ООН, входят сбор, обработка и публикация статистиче-



Рис. 2.1. Статистические организации ООН и другие международные организации

ских данных по международной статистике, получаемых от статистических служб государств — членов ООН.

Статистическое бюро готовит и выпускает целый ряд статистических изданий, в которых данные приводятся по странам, группам стран. К ним относятся:

- статистический ежегодник ООН «Statistical Yearbook». В нем освещаются вопросы населения, трудовых ресурсов, показатели развития отдельных отраслей, цены, макроэкономические, финансовые показатели, показатели культуры и т.д;
- ежемесячный статистический бюллетень «Monthly bulletin of Statistics»;
- демографический ежегодник «Demographic Yearbook». В нем размещаются сведения о территории, плотности населения, численности, движении, структуре, результаты национальных переписей населения, состояния населения. Информация представляется приблизительно по 220 странам и территориям. Наряду с официальной статистикой отдельных стран в ежегоднике помещены оценки Комиссии по народонаселению ЭКОСОС ООН;
- ежегодник по статистике международной торговли «Yearbook of International Trade Statistics», в котором приводятся цифры внешнеторгового оборота стран и торговых союзов по отдельным видам товаров, уровню цен и т.д.;
- ежегодник по статистике национальных счетов «Yearbook of National Accounts Statistics», где представлены данные о макроэкономических показателях производства, распределения и финансовой деятельности, по отдельным странам, секторам экономики, данные об инвестициях в основной капитал, потреблении, платежных балансах и т.д.

Рекомендации Статистической комиссии не являются обязательными для национальных статистических служб, однако при передаче статистических данных в ООН они должны представлять их в соответствии с действующими международными стандартами.

Внедрение в практику принятых Статистической комиссией решений возложено на Статистическое бюро, основная деятельность которого заключается в следующем: подготовке всех документов для сессий Статистической комиссии и ЭКОСОС; разработке проектов методологических руководств

и рекомендаций по стандартным показателям и их системам; сборе и анализе замечаний национальных служб по этим проектам для обобщения и т.д. Одобренные международные рекомендации и стандарты Статистическое бюро публикует в двух основных сериях документов ООН: методические записки (серия F) — как рекомендации ООН и статистические доклады (серия M) — как технические документы. В настоящее время разработано более 150 международных стандартов по статистике.

Основным стандартом отражения хозяйственной деятельности в стране являются международные рекомендации по системе национальных счетов (СНС).

Поскольку СНС основывается на принципах двойной записи, Статистическая комиссия принимает усилия по взаимной увязке принципов бухгалтерского и статистического учета на основе гармонизированной версии СНС. Для этих целей привлекаются международные профессиональные организации — Комитет по стандартам бухгалтерского учета и Международная федерация бухгалтеров. Их опыт разработки международных стандартов бухгалтерского учета обобщается и координируется Межправительственной рабочей группой экспертов ООН по международным стандартам учета и отчетности, которая помогает внедрению в практику стандартов, разрабатываемых Комитетом (к концу XX в. внедрено примерно 30 стандартов).

К статистическим стандартам можно отнести: международные стандарты классификаций, например классификации отраслей экономики, товаров, видов деятельности и т.д.; международные классификации показателей отраслевых статистик; стандарты по методологии исчисления статистических показателей и их систем и, наконец, международные справочники, такие, как «Справочник международных мер и весов», «Номенклатура географических регионов для нужд статистики», «Справочник пересчета данных о сельскохозяйственной продукции», «Название стран мира и национальностей», «Таможенные зоны в мире», «Указатель международных стандартов для статистики», «Перечень статистических показателей, разрабатываемых международными организациями», «Справочник международных статистических организаций» и т.д. На основе этих международных стандартов ООН рассылает

по всем странам мира специальные бланки-вопросники для сбора данных, которые затем публикуются в статистических изданиях ООН.

Большую работу по международной статистике выполняют региональные экономические комиссии ООН: Европейская экономическая комиссия (ЕЭК), Экономическая и социальная комиссия для Азии и района Тихого океана (ЭСКАТО), Экономическая комиссия стран Африки (ЭКА), Экономическая комиссия для Латинской Америки (ЭКЛА) и Экономическая комиссия ООН для Западной Азии (ЭКЗА). При каждой комиссии существуют статистические отделы.

Под руководством региональных экономических комиссий и Статистической комиссии ООН проводится большая статистическая работа. Так, ежегодно с 1953 г. в Женеве проводятся сессии Конференции европейских статистиков. В состав Конференции входят руководители центральных статистических управлений европейских стран или равные им по рангу должностные лица из стран, участвующих в работе Европейской экономической комиссии (США также являются членом этой комиссии). Конференция занята разработкой статистической методологии, стандартов, программ, является форумом для европейских консультаций в отношении этих стандартов. Подготовкой материалов для сессий занимаются рабочие группы, а также группы докладчиков из специалистов по отдельным отраслям статистики. Конференция координирует деятельность статистических служб ЕЭК по отдельным направлениям. Конференция проводится под руководством и в тесном контакте со Статистическим отделом Секретариата ЕЭК.

К основным статистическим публикациям ЕЭК можно отнести ежемесячное издание «Статистические показатели текущих изменений в странах Европы», два ежеквартальных, один полугодовой и 10 ежегодных бюллетеней по статистике производства и потребления в Европе энергии, стали, машин, жилья, транспорта.

В Региональной экономической и социальной комиссии для Азии и района Тихого океана Конференции азиатских статистиков работают на постоянной основе. В их работе принимают участие специалисты — статистики стран — членов ЭСКАТО, включая Россию, США, Англию и Францию.

Задачи и цели Конференции азиатских статистиков такие же, как у Конференции европейских статистиков.

Основные статистические публикации ЭСКАТО: «Статистический ежегодник для Азии и района Тихого океана», ежеквартальный «Статистический бюллетень для Азии и Тихого океана», «Статистика внешней торговли стран Азии и района Тихого океана», где приводятся основные статистические показатели развития региона.

Конференция африканских статистиков решает ряд специфических задач, свойственных развивающимся государствам, — организацию государственных статистических служб, внедрение рекомендуемых принципов сбора и обработки данных, развитие статистических методов анализа, подготовку специалистов. Экономическая комиссия стран Африки при помощи своих статистических служб издает «Статистический ежегодник», сборник «Статистика внешней торговли стран Африки» (серия А — «Направление торговли» и серия В — «Торговля отдельными товарами»), «Бюллетень статистической и экономической информации по Африке» (выходит 3 раза в год), ежегодник «Обзор экономического положения Африки». В работе Конференции африканских статистиков в качестве наблюдателей принимают участие представители России.

Экономическая комиссия стран Африки при помощи ООН и своих статистических служб проводит ряд исследований в различных областях современной статистики, в демографии, сельском хозяйстве, производстве.

Основную работу по координации статистической деятельности в Латинской Америке осуществляет Международный американский статистический институт (МАСИ), подчиненный Организации американских государств. Международный американский статистический институт является членом Международного статистического института и имеет консультативный статус при ЭКОСОС. Региональная экономическая комиссия для Латинской Америки не имеет региональной конференции статистиков, хотя, как и ЭКЗА, занимается сбором, анализом и публикацией статистических данных по Латинской Америке. Она издает полугодовой «Статистический бюллетень по Латинской Америке», в котором приводятся данные по населению, трудовым ресурсам, внешней торгов-

ле, производству отдельных видов товаров как дополнение к «Обзору экономического положения Латинской Америки».

Остановимся на статистической деятельности специализированных учреждений ООН.

Международная организация труда (постоянно действующий орган — Международное бюро труда), которая имеет в своем составе статистический отдел, регулярно созывает международные конференции статистиков по вопросам труда и социального страхования.

Основные издания МОТ: «Ежегодник статистики по труду», охватывающий более 180 стран за 10 лет, а также «Бюллетень статистики по труду», издаваемый ежеквартально, в котором, кроме годовых данных за 10 лет, приводятся данные по кварталам и месяцам за последние 3 года; «Дополнение к бюллетеню», выходит 8 раз в год и содержит сведения о занятости по отдельным отраслям, условиям труда, индексам потребительских цен. Кроме того, в каждом июньском выпуске содержатся данные за предшествующий октябрь по ежегодному обследованию заработной платы 41 профессии, продолжительности рабочего дня и уровню цен по 41 виду потребительских товаров. Приводятся статистика доходов и расходов домашних хозяйств и расходы на социальное обеспечение.

Работа Продовольственной и сельскохозяйственной организации ООН (ФАО) в области статистики очень обширна. Ряд ее подразделений имеет статистические отделы. ФАО публикует «Ежегодник по производству», «Ежегодник по торговле», «Ежегодный обзор положения в области удобрений», «Ежегодный бюллетень экономики и статистики сельского хозяйства», «Средний трехгодичный продовольственный баланс», «Ежегодник лесопромышленной продукции», «Ежегодник по вопросам рыболовства», «Ежегодник статистики мировой торговли зерном» и многие другие издания. В них содержится информация о производстве, торговле, ценах, потреблении, средствах производства, структуре сельскохозяйственной деятельности и т.д.

Статистические отделы ЮНЕСКО заняты сбором, обработкой, публикацией и анализом статистических данных по вопросам образования, науки и техники, культуры, информации и т.д. Основными публикациями являются: «Статистический ежегодник ЮНЕСКО», в котором приводятся данные о

населении, образовании, науке и технике, библиотеках и музеях, книжных изданиях, потреблении бумаги и статистике кино, радио и телевидения; «Краткий статистический обзор образования в мире», издаваемый один раз в два года; мировые обзоры и статистические доклады: «Мировой обзор образования», «Мировые средства массовой информации», «Статистика специального образования», «Получение образования за границей» и др.

Международная организация гражданской авиации (ИКАО) собирает и публикует данные о коммерческом воздушном транспорте, объеме перевозок, воздушном парке и персонале, финансовые показатели. Издает «Статистический сборник ИКАО» ежеквартально, а более подробные публикации включают следующие серии: «Перевозка авиалиний», «Воздушное движение», «Воздушный флот и персонал», «Финансовые сведения», «Нерегулярный воздушный транспорт», «Зарегистрированные самолеты гражданской авиации», «Авиадвижение в аэропортах».

Вся статистическая информация Всемирной организации здравоохранения разрабатывается по следующим направлениям: развитие служб здравоохранения, методология статистики здравоохранения и международной классификации болезней. Основная статистическая работа проводится в региональных отделениях. Всемирная организация здравоохранения ежемесячно публикует «Доклад по мировой санитарной статистике», в первой части которого приведены показатели смертности, заболеваемости инфекционными болезнями, вакцинации, обеспеченности медицинскими услугами и т.д., а во второй — причины смертности и заболеваемости, расходы на медицинское обслуживание и т.д. Кроме этого публикуется «Ежегодник мировой санитарной статистики», где, помимо специальных данных, приводятся данные о численности населения, его динамике, структуре, естественном движении, распределении по месту жительства.

Международный банк реконструкции и развития (МБРР), Международная ассоциация развития (МАР) и Международная финансовая корпорация (МФК) являются самостоятельными юридическими лицами, однако их деятельность тесно связана друг с другом, что проявляется в работе статистических служб. Статистическая работа, выполняемая МБРР, основыва-

ется на материалах служб МАР и МФК. Их основная задача — предоставление необходимой информации по странам, регионам и проектам. Банк является в основном потребителем, а не источником информации. Единственной группой первичных данных, публикуемых банком, являются данные о внешней задолженности, потоках капитала, погашении долга, национальных счетах и платежных балансах. Банк ежегодно организует до 70 экспедиций в развивающиеся страны с целью сбора информации и получения оперативных оценок состояния экономики. Регулярно публикуются: «Годовой отчет Международного банка и Международной ассоциации развития» с приложениями, в котором дается подробная информация о международной задолженности; ежегодник «Тенденция в развивающихся странах», содержащий данные о народонаселении, социально-экономическом росте, движении международного капитала, внешней задолженности и международной торговле.

Международный валютный фонд (МВФ) подготавливает для внутреннего пользования и публикаций текущие данные о международных валютных запасах, финансовых и банковских операциях, государственных финансах, процентных ставках, ценах, заработной плате, производстве, международной торговле и национальных счетах. МВФ публикует сборник «Международная финансовая статистика» («International Financial Statistics») и приложение к нему, где рассматриваются вопросы внутренних и международных финансов, содержатся данные более чем по 100 странам о внутренних валютных проблемах и проблемах платежного баланса, валютных курсах, валютных резервах международной торговли и т.д.; издает ежемесячник «Направления торговли» и «Ежегодник платежных балансов».

Всемирный почтовый союз издает ежегодник «Статистика почтовых служб», содержащий сведения по странам и территориям о почтовых отправлениях, финансовых операциях почт, данные о наличии почтовых учреждений и их технической оснащенности.

Международный союз электросвязи публикует «Статистический ежегодник общих поставщиков электроэнергии», «Статистику электросвязи».

В Европе в настоящее время действуют три крупных центра международной статистики, занятые сбором и обработкой

статистической информации: отдел статистики Европейской экономической комиссии ООН; Статистическое бюро Организации экономического сотрудничества и развития (ОЭСР) и Европейская статистическая комиссия (Евростат). В «Основных принципах официальной статистики в регионе Европейской экономической комиссии», принятых на 47-й сессии ЕЭК 15.04.1992 г., подчеркиваются значимость официальной статистической информации, необходимость обеспечения ее точности и объективности, соответствия международным стандартам и принципам профессиональной этики (полностью этот документ приведен в приложении 2).

Основными задачами Евростата являются: подготовка и предоставление статистической информации для руководства Европейского экономического, валютного и политического союза (ЕС); постоянное совершенствование статистической информации; помощь национальным статистическим службам в унификации подходов к организации статистических наблюдений в странах ЕС и в единой методологии расчета показателей. Предоставление информации в Евростат — обязанность национальных статистических служб.

Евростат включает следующие подразделения: распространения статистической информации и компьютеризации экономической статистики и государственного бюджета, международной торговой статистики и внешних сношений, статистики предприятий, социальной и региональной статистики, сельскохозяйственной статистики и статистики окружающей среды. Использование единых международных стандартов для системы национальных счетов, а также разработка методологии расчетов паритета покупательной способности валют (ППС) стран ЕС в значительной степени облегчают работу по согласованию между странами экономических программ и взаимных расчетов. В рамках Евростата работает Общественный совет, включающий представителей всех стран ЕС и различных потребителей информации, с учетом интересов которых и разрабатывается план работ Евростата.

Деятельность Евростата регламентируется законодательно, его функции четко определены и разграничены.

Большую статистическую работу ведут Организация экономического сотрудничества и развития — ОЭСР (OECD) и ее многочисленные отраслевые комитеты, которые постоян-

но готовят статистические материалы и публикуют их в разнообразных изданиях, в частности «OECD Outlook Express», «Main Economical Indicators» и др.

Координация деятельности статистических служб стран — членов СНГ осуществляется созданным в 1992 г. Статистическим комитетом Содружества Независимых Государств. Публикуются статистические сборники по странам СНГ и другим государствам ближнего зарубежья.

Статистические публикации — это один из возможных источников статистической информации. Используя его, следует критически относиться к статистическим данным, прикидывая, насколько та или иная цифра реальна. Полезно иметь данные из разных источников. Наверняка они не совпадут, но доверие будет к тем данным, которые имеют близкие значения. Пользоваться лучше теми данными, способ получения которых понятен. Достоверность данных государственной статистики определяется тем, что это результат профессиональной деятельности специально подготовленных работников, использующих единую методологию, соответствующую в большинстве случаев международным стандартам, дающую возможность проследить динамику какого-либо показателя за ряд лет.

Если соответствующих данных нет в статистических сборниках, то можно получить их самим, т.е. провести *статистическое наблюдение* — научно организованный сбор данных. В системе государственной статистики не менее трети всего объема работ связано с получением данных.

Кем бы и когда бы ни проводилось статистическое наблюдение, оно должно быть организовано по определенным правилам, соблюдение которых позволяет обеспечить надежную основу статистического исследования.

2.3. Требования, предъявляемые к собираемым данным. Формы организации и виды статистического наблюдения

Собираемые данные должны отвечать двум требованиям: достоверности и сопоставимости. *Достоверность* — это соответствие данных тому, что есть на самом деле. Вся методика,

организация и техника проведения статистического наблюдения должны быть нацелены на обеспечение достоверных данных. Для того чтобы понять характер задач, возникающих при этом, представим статистическое наблюдение в виде взаимодействующих компонентов (рис. 2.2).



Рис. 2.2. Составляющие статистического наблюдения

Очевидно, что достоверность данных зависит как от характеристик самого статистика — его профессиональной подготовки, коммуникабельности, организационных навыков и т.д., так и от качества используемого инструментария — программы наблюдения, бланков, анкет, инструкций по их заполнению. Они в конечном счете тоже зависят от статистика. На достоверность данных влияет и подготовленность объекта к статистическому обследованию. Это может быть сделано в форме предварительного извещения населения о предстоящем обследовании — в газетах, по радио, телевидению (как это делалось, например, перед началом Всероссийской переписи населения 2002 г.). Влияет на достоверность и упорядочение названия улиц и нумерации домов, квартир и т.д.

На достоверности данных сказывается социальная функция показателя. Известно, например, о фактах недостоверности данных о младенческой смертности (смертности детей до одного года). Основной недоучет составляют случаи, когда факт рождения ребенка, умершего вскоре после рождения, умышленно регистрируется как мертворождение, а часть случаев мертворождений записывается как поздние выкидыши и не регистрируется в органах ЗАГСа. Недостоверными могут быть данные о характере и числе преступлений, профессиональной заболеваемости и т.п., т.е. те данные, которые сигнализируют о «здоровье» общества.

Условиями обеспечения достоверности являются полнота охвата наблюдаемого объекта; полнота и точность регистрации данных по каждой единице наблюдения.

Чтобы данные об отдельных явлениях можно было обобщать, они должны быть *сопоставимы* друг с другом: собирать-

ся в одно и то же время по единой методике. Кроме того, должна быть обеспечена сравнимость с прошлыми исследованиями, чтобы можно было понять, как изменяется явление.

Сравнимость данных разных наблюдений выполняется, если использовались одно и то же определение единицы наблюдения, одна и та же методика регистрации первичных признаков и методика расчета вторичных признаков, таких, как себестоимость, производительность труда, рентабельность, ликвидность и т.д.

Важным условием сравнимости является сохранение времени проведения наблюдения и периода или момента, к которому относятся регистрируемые данные. Например, численность студентов университета определяется на начало учебного года, стипендиальный фонд — на полгода (или год). Обычно рекомендуется, чтобы данные соответствовали хотя бы одному полному циклу изучаемого процесса, например учебному, хозяйственному или финансовому году. Если сильно влияет сезонность, данные должны собираться по месяцам или по кварталам. Время наблюдения выбирается таким образом, чтобы наблюдаемый объект находился в наиболее стабильном состоянии.

Статистическое наблюдение подразделяется на виды — по времени наблюдения и по охвату единиц наблюдения.

По времени регистрации фактов различают непрерывное (текущее), периодическое и единовременное наблюдение. *Непрерывное* (*текущее*) *наблюдение* ведется систематически, постоянно, непрерывно, по мере возникновения явлений. Например, регистрируются в ЗАГСе рождения и смерти, браки и разводы, на предприятиях учитываются выпуск продукции, явки и неявки работников, расчеты с дебиторами и кредиторами, поступление денег в кассу и денежные выплаты и т.п. При *периодическом наблюдении* регистрация проводится через определенные (обычно одинаковые) промежутки времени, например учет успеваемости студентов по данным экзаменационных сессий. *Единовременное наблюдение* проводится один раз для решения какой-либо задачи или повторяется через неопределенные промежутки времени по мере надобности, например перепись жилого фонда, школьная перепись, перепись скота, плодово-ягодных насаждений и т.д.

Применение на практике того или иного вида наблюдения зависит от специфики исследуемого объекта. Так, функционирование общественного производства носит непрерывный характер: ежедневно производится и потребляется множество различных видов продукции, изменяются их запасы и т.д. Обеспечение бесперебойного производства требует непрерывного поступления сырья и материалов и их учета, систематического учета затрат на производство и его результатов. Иной характер носят изменения в составе населения по социальному или национальному признаку, образованию и пр. В обычных условиях для больших групп населения эти признаки несущественно изменяются в короткие промежутки времени, поэтому нет надобности в непрерывной их регистрации. Достаточно проводить переписи населения один раз в 10 лет.

Бывает, что для изучения одного и того же процесса используется как текущее, так и единовременное наблюдение. Например, потребление населения изучается государственной статистикой по данным текущего наблюдения (бюджетные обследования). В то же время многими исследовательскими коллективами потребление изучается по данным единовременных наблюдений: фиксируются «обычные» дневные покупки продовольствия, иногда эти данные дополняются данными фактических покупок за последние 2—3 дня; фиксируются наличие предметов длительного пользования, покупки непродовольственных товаров за последний месяц, квартал или полгода и т.д.

По охвату единиц совокупности различают сплошное и несплошное наблюдение.

При *сплошном наблюдении* регистрации подлежат все без исключения единицы совокупности. Оно применяется, например, при переписи населения, сборе данных в форме отчетности, охватывающей предприятия разных форм собственности, учреждения и организации и т.д.

Развитие многоукладной экономики увеличило число объектов экономической деятельности. Это способствовало расширению практики *несплошного наблюдения*, которое, в свою очередь, подразделяется на способ основного массива, выборочное и монографическое.

При способе *основного массива* обследованию подвергается та часть единиц, которая вносит наибольший вклад в изу-

чаемую совокупность. Остальные, которые не играют большой роли в характеристике совокупности, исключаются из наблюдения, т.е. при этом методе отбираются и обследуются наиболее крупные единицы. Логика метода состоит в том, что крупные единицы могут практически определять интересующие нас статистические показатели. Например, вследствие концентрации производства в отрасли несколько наиболее крупных предприятий могут давать основной объем продукции, в то время как большая масса мелких предприятий выпускает ее незначительную часть. Это бывает при высоком уровне монополизма в отрасли экономики, особенно в условиях региона. Так, в Санкт-Петербурге в 1991 г. всего лишь на 7 предприятиях машиностроения и металлообработки, которые составляли 1,3% числа промышленных предприятий города, работало около 20% работников. На каждом из этих предприятий было занято свыше 10 тыс. человек, в эту группу входили такие гиганты, как «Кировский завод» — 25 тыс. человек, «Ленинец» — 22,9 тыс. человек и т.д. В подобных условиях логично наблюдать только наиболее крупные предприятия, а мелкие либо вообще игнорировать, либо провести досчет приходящейся на них доли продукции. Поскольку их доля невелика, ошибка при распространении данных основного массива на всю совокупность будет незначительной. Точность досчета зависит от того, какими сведениями о не охваченной наблюдением части совокупности мы располагаем.

Применение метода основного массива часто требует установления *ценза* — значения признака, которое ограничивает объект наблюдения. Например, обследуются предприятия с числом работников 500 человек и более или устанавливается, что обследованию подлежат малые предприятия с численностью работников до 100 человек (или до 200 человек). Такой метод называется цензовым.

Следует иметь в виду, что термин «ценз» употребляется в статистике не только в смысле пограничного значения признака, но и для обозначения переписей. В США, Англии цензами называют переписи населения, промышленности и т.д.

При *выборочном наблюдении* обследованию подвергается отобранная в определенном порядке часть единиц совокупности, а получаемые результаты распространяются на всю совокупность.

В выборке полностью реализуется основная идея несплошного наблюдения: получить информацию о всей совокупности, изучив лишь ее часть. Для того чтобы понять, хорошее пиво или плохое, не обязательно выпить целую бочку, то же можно сказать в отношении проверки качества любой продукции. В решении такого рода задач, да и во многих других случаях может помочь только выборка.

Выборочный метод играет все большую роль в отечественной статистике.

Обследования основного массива и выборки — это *массовые наблюдения*, охватывающие множество единиц. При *монографическом наблюдении* подробно описываются отдельные единицы совокупности в целях их углубленного изучения, которое не может быть столь же детальным при массовом наблюдении. Первоочередное внимание уделяется качественным сторонам явления, его поведению, ориентации, перспективе развития и т.д. Примерами монографических обследований могут служить этнографические обследования, когда изучается образ жизни семьи или нескольких семей и др.

В любом обследовании *источником* получения первичных данных могут быть непосредственное наблюдение, документы и опрос.

Непосредственное наблюдение осуществляется путем регистрации изучаемых единиц и их признаков на основе непосредственного осмотра, подсчета, взвешивания, показаний приборов и т.д. Так, во время переписи вагонов проводится осмотр каждого вагона. Примером непосредственного наблюдения являются: регистрация цен и объема реализации товаров на рынках; метеорологические наблюдения — регистрация температуры воздуха, снежного покрова, суммы осадков; инвентаризация остатков товарно-материальных ценностей на складе.

Документальный способ наблюдения основан на использовании в качестве источника статистических сведений различных документов первичного учета предприятий, учреждений и организаций, поэтому этот способ наблюдения часто называют отчетным. Он применяется, например, при переоценках основных фондов (средств) предприятий и организаций, которые составляют основу начисления амортизации, анализа использования фондов и их структуры, особенно в условиях

инфляции. При заполнении государственной статистической отчетности по переоценке каждым самостоятельным предприятием любой отрасли и формы собственности используются следующие данные первичной учетной информации: инвентаризационные описи, инвентарные карточки основных фондов, технические паспорта или другая соответствующая документация и данные бухгалтерского учета.

Непосредственное наблюдение и документальный способ обеспечивают наибольшую достоверность статистических данных.

При *опросе* источником данных являются сведения, которые дают опрашиваемые лица. При этом могут быть использованы разные *способы сбора данных*: экспедиционный, корреспондентский и саморегистрация.

Экспедиционный способ заключается в том, что специально подготовленные регистраторы на основе опроса заполняют переписные формуляры, одновременно контролируя правильность получаемых ответов. Этот способ обеспечивает достаточно точные результаты, но он дорогостоящий. В отечественной статистике экспедиционный способ используется при переписях населения.

Корреспондентский способ заключается в том, что статистические или другие организации рассылают специально разработанные бланки и инструкции к их заполнению отдельным организациям или специально подобранным лицам, давшим согласие периодически заполнять бланки и присылать статистическому органу в установленные сроки. Например, Научно-исследовательский институт по изучению спроса населения на товары народного потребления и конъюнктуры торговли создал сеть корреспондентов в каждом регионе, которые периодически сообщают в центр сведения о покупательском спросе населения, товарном обеспечении в данной местности и другую информацию. Преимуществом этого способа является его дешевизна, однако он не всегда обеспечивает хорошее качество сведений, так как зависит от уровня восприятия вопросов опрашиваемым, от его ответственности — отправит он заполненную анкету или нет.

При *саморегистрации,* или самоисчислении, работники той организации, которая проводит опрос, раздают опросные листы или анкеты опрашиваемым лицам, инструктируют их,

а затем собирают заполненные формуляры, контролируя полноту и правильность полученных сведений. Этот способ используется в государственной статистике при бюджетных обследованиях семей, проведении некоторых переписей и т.д. В последние годы при сборе статистической информации начинают использовать безбумажные технологии.

Заметим, что при любом методе проведения статистическое наблюдение пассивно: статистика хочет как можно точнее зарегистрировать данные без какого-либо влияния на наблюдаемый процесс. Принципиально иным методом сбора данных является *эксперимент*. В этом случае статистику принадлежит активная роль: он должен не только наблюдать, но и полностью контролировать ситуацию, планировать эксперимент и реализовать свой план. Эксперимент позволяет выявить влияние каких-либо установленных ограничений или нагрузок на поведение людей. Например, влияние на скорость реакций человека пребывания без сна в течение одних, двух, трех суток. Эксперимент традиционно входил в круг методов биологической, медицинской статистики, приложений статистического метода в естественных науках. В настоящее время все большее распространение получают идеи «социального эксперимента».

2.4. Подготовка статистического наблюдения

Для того чтобы провести статистическое наблюдение, нужно сформулировать его цель и основные гипотезы, которые должны быть проверены по данным наблюдения. Эта стадия работы определяет последующие, поэтому обычно все решения вырабатываются коллективно в ходе обсуждения проблем предстоящего наблюдения. На этой стадии дается определение объекта и единицы наблюдения, разрабатывается и утверждается программа наблюдения, а также сроки проведения, источники и способы сбора данных, состав исполнителей.

Определение *объекта наблюдения* включает определение единицы наблюдения, территории и времени наблюдения.

Единица наблюдения — это то явление, признаки которого подлежат регистрации. Совокупность единиц наблюдения составляет *объект наблюдения*. Как уже отмечалось, для определения границ объекта наблюдения нередко устанавливается

ценз — значение признака (или нескольких признаков), позволяющее отделить единицы наблюдения от других явлений. В самом деле, трудно установить границы даже, казалось бы, очевидного объекта — совокупности промышленных предприятий: что входит в понятие «промышленное предприятие», а что нет. Входят ли в круг промышленных предприятий предприятия по ремонту и мойке автомобилей, закупке и переработке фруктов и т.д.? Устанавливать ли цензовые значения только по численности работников или по стоимости производственного оборудования? При проведении переписи населения возникают вопросы: учитывать ли тех граждан, которые длительное время работают за границей? как учитывать тех, кто находится в заключении, на службе в армии? и т.д. Все эти вопросы требуют всестороннего обсуждения. Их решение основано на том, что является конечным результатом, что должно быть получено в результате исследования. Если не предусмотреть чего-то на начальной стадии, это скажется на качестве всего исследования.

Территория проведения наблюдения охватывает все места нахождения единиц наблюдения; ее границы зависят от определения единицы наблюдения.

Время наблюдения — это то время, к которому относятся собираемые данные. Время регистрации данных для всех единиц устанавливается единое — для предупреждения неполного учета или повторного счета, а также для обеспечения сопоставимости данных.

При изучении объектов наблюдения, численность и характеристика которых непрерывно изменяются, устанавливается *критическая дата*, по состоянию на которую собираются сведения. При переписях обычно устанавливают время начала и окончания регистрации данных. Так, последняя Всероссийская перепись населения проводилась в течение 8 дней — с 9 по 16 октября 2002 г.; 5%-ная микроперепись населения РФ проводилась в течение 10 дней — с 14 по 23 февраля 1994 г. И в том, и в другом случае время наблюдения приходилось на период и даты, когда у работающих меньше отпусков, нет государственных праздников или каникул у школьников и студентов.

При изучении такого подвижного объекта, как население, недостаточно установить время наблюдения — ведь состав населения России и его характеристики постоянно меняются:

в среднем каждую минуту в нашей стране рождаются 3 человека и умирают 3—4 человека. Поэтому данные регистрируются по состоянию на определенный момент времени, называемый *критическим моментом наблюдения*. В качестве критического момента во Всероссийской переписи населения, проведенной 9—16 октября 2002 г., было принято 0 часов с 8 на 9 октября. Соответственно в бланки переписи заносились все живущие на данный момент и не вносились родившиеся после 0 часов с 8 на 9 октября 2002 г. и умершие до этого времени.

При переоценке основных фондов устанавливается *критическая дата*, по состоянию на которую учитываются основные фонды (здания, сооружения, оборудование, транспорт и т.д.). Например, одна из переоценок проводилась по состоянию на 1 января 1994 г. Все предприятия, владевшие основными фондами на эту дату, должны были показать сведения о них в отчете; если в период между 1 января 1994 г. и моментом заполнения бланка отчетности какие-либо фонды были проданы, переданы другому владельцу, то новый владелец не включал их в свой отчет во избежание двойного учета.

Определение объекта наблюдения, его территориального размещения важно для установления объема работ, который нужно выполнить в период наблюдения. Если наблюдение планируется провести в форме отчетности, то составляется список подотчетных предприятий и организаций. При специально организованном наблюдении определение объема работ необходимо для расчета численности работников, требуемых для выполнения обследования в установленные сроки. Рассчитывается дневная норма работы одного регистратора (счетчика) с учетом сложности программы наблюдения, трудоемкости заполнения формуляра наблюдения и размещения объекта. В сельской местности, например, где плотность застройки намного ниже городской, дневная норма устанавливается меньше, чем в городах. В целях лучшей организации наблюдения и контроля за качеством материала вся территория разбивается на отдельные счетные участки; 20—30 счетных участков при переписи населения образуют инструкторский участок, руководимый инструктором.

Проведение массовых работ требует участия множества исполнителей (в переписях населения участвуют тысячи счетчиков). Все они должны пройти специальное обучение — ин-

структаж и провести пробное заполнение тех формуляров, которые предполагается использовать в статистическом наблюдении. Должна быть составлена смета на проведение специального обследования, в которой предусматриваются размножение материалов наблюдения (бланков, инструкций), оплата услуг средств связи, транспорта, работа инструкторов, счетчиков и др. Статистическое обследование — дорогостоящая и трудоемкая процедура. Проведение обследований должно быть обосновано и подкреплено финансовыми, материальными и трудовыми ресурсами.

Программа наблюдения включает признаки, подлежащие регистрации по каждой единице наблюдения. Ее содержание зависит от целей и задач обследования. В какой-то мере программа наблюдения зависит и от выделенных средств: мало средств — программа может быть короче, или число наблюдаемых единиц меньше. Поэтому первый принцип составления программы наблюдения — никаких сведений, не относящихся к данному обследованию («на всякий случай»).

Второй принцип, немаловажный для получения достоверных данных при опросах, — не включать в программу наблюдения те вопросы, которые могут показаться людям подозрительными и на которые можно заведомо ожидать неточных ответов. Например, при изучении потенциальной эмиграции не стоит включать в анкету прямой вопрос типа: «Собираетесь ли вы уехать за границу на длительное время или навсегда?». Более эффективно использовать систему вопросов, составленных таким образом, чтобы их сочетание позволяло сделать те заключения, которые бы вы хотели получить с помощью ответов на прямой вопрос. Или, понимая, что точную сумму доходов и сбережений состоятельные люди скорее всего не укажут, имеет смысл задать косвенные вопросы, например: «Есть ли среди ваших знакомых люди с месячным доходом 10 тыс. долл. и выше?» и т.д. Не рекомендуется задавать вопрос: «Сколько денег вы заработали в прошлом году?», лучше спросить: «Какая из следующих категорий соответствует вашему доходу в прошлом году:

до 100 тыс. руб.
100—150 тыс. руб.
150—200 тыс. руб.
200—250 тыс. руб.

250—300 тыс. руб.
300 тыс. руб. и более».

Следует помнить, что ответ зависит от формы, в которую облечен вопрос. Например, в анкете имеется вопрос: «Вы согласны с тем, что высокое качество школ, больниц, общественных услуг напрямую зависит от повышения налогов?». Делая акцент на качество общественных учреждений, вы скорее получите положительный ответ, чем в том случае, если спросите: «Вы сторонник повышения налогов в следующем году?».

С целью уточнения формулировок вопросов, определения того, как они «работают», проводят пробные, или пилотные, обследования. Например, при подготовке к Всероссийской переписи населения 2002 г. были проведены две пробные переписи — в 1997 и 2001 гг., по результатам которых проводилась корректировка вопросов переписного листа.

Программа наблюдения всегда включает опознавательные признаки; вопросы, непосредственно связанные с целью исследования; контрольные вопросы. Выделение последних весьма условно, поскольку один и тот же вопрос может выполнять как содержательную, так и контрольную функцию. Например, программа переписи населения содержит вопросы о возрасте, образовании, семейном положении, наличии детей, их возрасте, образовании и т.д. Все они логически связаны, что позволяет контролировать правильность ответов. Те же принципы лежат в основе бюджетных обследований — вопросы о доходах и расходах выполняют и познавательную функцию, и функцию взаимного контроля.

Опознавательные признаки позволяют идентифицировать единицу совокупности, к которой относятся регистрируемые данные. В социологических обследованиях вопрос обычно анонимный. Однако чтобы избежать недоучета и повторного счета, каждой единице наблюдения (опрашиваемому) присваивается какой-либо номер (шифр), а также фиксируется место проживания (населенный пункт). При сборе данных в форме отчетности опознавательными признаками являются название предприятия (организации), его шифр в регистре государственной статистики, отраслевая принадлежность, адрес, номер телефона, факса и т.д.

Все вопросы программы наблюдения ориентированы на определенную *форму ответа*: цифровую, альтернативную («да»

или «нет»), многовариантную, когда ответ состоит в выборе одного или нескольких вариантов из множества предлагаемых. Так, на вопрос о возрасте ответ дается в количественной форме — указывается число исполнившихся лет; то же — на вопрос о стаже работы; ответ на вопрос о наличии автомобиля или дачи будет в альтернативной форме — «да» или «нет»; ответ на вопрос о степени удовлетворенности работой или учебой выбирается из предлагаемого меню. Обычно такое меню строится по принципу симметрии: абсолютно негативное (или, наоборот, абсолютно позитивное) отношение, затем — более мягкая оценка, затем — выражение полной индифферентности, после чего оценки переходят в противоположную область: если были негативные, то теперь — позитивные и наоборот.

Предлагаемые варианты ответов называются *подсказом*. Наличие подсказа обеспечивает единообразное понимание вопросов и облегчает последующую обработку данных, так как каждый предлагаемый вариант ответа имеет свой код или шифр и работа по обработке ведется лишь по тем вариантам ответов, которые не были предусмотрены в подсказе и вписывались самими опрашиваемыми (*респондентами*).

Приведем в качестве примера фрагмент из анкеты читателей молодежной газеты «Смена».

Как к вам попал этот номер «Смены»?

001 — подписчиком газеты являюсь лично я;

002 — взял у знакомых;

003 — купил в газетном киоске;

004 — газету выписывают у меня дома;

005 — другой ответ.

Наличие кодов облегчает обработку собранного материала, которая начинается сразу же, как только статистик убедился, что получены данные от всех единиц и даны ответы на все вопросы.

В переписях населения и других специальных обследованиях, проводимых государственной статистикой, подсказы обычно включают все варианты ответов (без дописывания). Например, вопрос о типе жилого помещения в программе переписи 2002 г. включал варианты ответов: индивидуальный дом, отдельная квартира, общая (коммунальная) квартира, общежитие, другое жилое помещение, снимает жилое помещение.

Составление программы наблюдения — сложная и ответственная задача. В государственной статистике разработкой программы специальных обследований занимаются специалисты Госкомстата России и НИИ при участии представителей Научно-методологического совета и заинтересованных организаций. Программы таких важных и массовых работ, как перепись населения, переоценка основных фондов и других, обсуждаются на специальных совещаниях, в печати, что обеспечивает их высокое качество.

Инструментарий статистического наблюдения включает *формуляры* и *инструкции* по их заполнению. Формуляры наблюдения — это бланки, опросные листы, анкеты и т.д., на которых напечатаны вопросы программы наблюдения; в них затем заносятся собираемые сведения. Соответственно в формуляре должно быть предусмотрено место для вопроса и ответа. Обычно в верхней части формуляра или на первой странице располагаются опознавательные признаки, слева — вопросы программы наблюдения, справа — место для ответов. Формуляр наблюдения может быть *карточным* (индивидуальным) или *списочным*. В первом случае он предназначен для записи данных только по одной единице наблюдения, во втором — по нескольким. В переписи населения РФ 2002 г. была принята списочная форма — формуляр заполнялся на домохозяйство. При этом если число членов домохозяйства превышало 5 человек, то использовался дополнительный бланк, а в опознавательной части проставлялись буквенные обозначения бланка (*а*, *б* и т.д.).

Качество данных статистического наблюдения зависит не только от перечисленных факторов, но и от подготовленности счетчиков (регистраторов, интервьюеров). Для них организуется инструктаж по разъяснению вопросов анкеты (или другого формуляра наблюдения) и пользованию инструкцией. Объясняется, например, что при наличии подсказов счетчик обязан ознакомить респондента со всеми вариантами ответов, не выделяя из них те, которые он сам считает наиболее вероятными. Затем проводится пробное заполнение анкет, итоги которого коллективно обсуждаются.

Доброжелательность счетчика, его умение вступать в контакт с людьми влияют на атмосферу опроса, а значит, и на его результаты. Важной этической проблемой является аноним-

ность данных опроса. Уверенность в анонимности снимает напряженность при регистрации мнений, суждений, пожеланий, а также характеристики благосостояния (чем владеет респондент, имеет ли сбережения, что из «крупных» вещей приобрел за последний год и т.д.). Иногда в интересах планирования наблюдения и контроля данных полной анонимности респондентов нет, но конфиденциальность информации обеспечивается. Так, если для проведения опроса с целью изучения уровня бедности в России в качестве основы выборки использовались списки избирателей, то соответствующий код респондента позволяет идентифицировать его. В таких случаях респондент должен быть убежден, что его ответы как персональные никогда не будут использованы. Они войдут в общую совокупность ответов и послужат основой расчета обобщающих показателей.

Как бы тщательно ни была составлена программа наблюдения и разработан формуляр, для обеспечения единообразия его заполнения, толкования вопросов все же необходима инструкция. Этот документ содержит объяснения вопросов программы с конкретными примерами, указания по взаимосвязи вопросов. Инструкция издается либо в виде отдельной брошюры, либо дается в подсказах, либо на самом формуляре наблюдения (обычно на оборотной стороне). Сфера специальных обследований непрерывно расширяется, и от их качества во многом зависит, увеличится или уменьшится число лиц, скептически относящихся к статистике.

2.5. Статистическая отчетность

Статистическая отчетность — особая форма организации сбора данных, присущая только государственной статистике. Она проводится в соответствии с федеральной программой статистических работ. Государственная статистика использует все виды статистических наблюдений (регулярную отчетность, единовременные учеты, различного рода переписи, выборочные, анкетные, социологические, монографические обследования и т.д.), формы и программы которых утверждены Государственным комитетом Российской Федерации по статистике или по согласованию с ним органами государственной статистики в составе Российской Федерации

(краев, областей, автономной области и автономных округов, городов Москвы и Санкт-Петербурга).

Сведения о деятельности предприятий, организаций поступают в статистические органы в установленные сроки в виде определенных документов (отчетов). Бланки таких отчетов называют *формами статистической отчетности.* Каждая из них имеет свой шифр и название.

Программа отчетности, т.е. перечень собираемых сведений, методика их определения и форма бланка отчетности, разрабатывается и утверждается Госкомстатом России. Формы отчетности, включающие финансовые результаты, утверждаются, кроме того, и Минфином РФ.

Отчетность различается *по периодичности.* Она бывает срочная — содержит данные за месяц и менее (декаду, сутки), а также квартальная; полугодовая; годовая. Наиболее подробной является программа годовой отчетности.

Статистическое наблюдение в форме отчетности использует только один источник данных — документы. Прежде всего это документы бухгалтерского учета предприятий, организаций. Госкомстат России проводит политику *унификации* отчетности предприятий разных отраслей экономики.

Предприятия и организации любых форм собственности обязаны представлять отчетность в установленные сроки по утвержденной форме. Нарушением сроков представления государственной статистической отчетности считается опоздание на одни сутки, а опоздание более чем на одни сутки рассматривается как непредставление отчетности. Искажением отчетных данных считается неправильное их отражение в государственной статистической отчетности, допущенное как в результате умышленных действий должностных лиц с целью сокрытия доходов и в других корыстных целях, так и вследствие нарушения действующих инструкций и методологических указаний по составлению статистической отчетности, а также арифметических ошибок.

Важной функцией государственной статистики является определение *круга подотчетных единиц.* С этой целью все предприятия, организации, объединения независимо от формы собственности, а также граждане, занимающиеся предпринимательской деятельностью, представляют в органы государственной статистики учредительные документы для

присвоения идентификационных кодов, определения классификационных признаков на основе общероссийских классификаторов технико-экономической информации для включения в Единый государственный регистр предприятий и организаций всех форм собственности и хозяйствования (ЕГРПО) и отражения в государственной статистической отчетности.

При реорганизации или ликвидации предприятия, учреждения, организации, объединения представляют органам статистики государственную статистическую отчетность за период своей деятельности в отчетном году до момента ликвидации на бланках форм годовой отчетности, а также нормативные акты о своей реорганизации или ликвидации для внесения изменений в ЕГРПО.

Отчетность дает необходимую информацию для государственных органов управления. Данные отчетности позволяют следить за динамикой объема промышленного производства и продукции других отраслей народного хозяйства, оценивать комплексность развития страны и регионов, изучать соотношения разных форм собственности по отраслям и регионам и сравнивать эффективность деятельности государственных и негосударственных предприятий и организаций.

Большое значение имеют стабильность отчетности, содержание ее форм. Только при этом условии обеспечивается возможность построения протяженных рядов динамики, а значит, выявления тенденций, анализа колеблемости, разработки прогнозов. Конечно, содержание отчетности — перечень форм, показателей — меняется со временем, но прежде чем внести какое-либо изменение, нужно решить, является ли оно действительно необходимым. Ведь отчетность подготавливают десятки тысяч работников бухгалтерских и финансовых отделов предприятий и организаций. Очевидно, что такая массовая форма сбора данных может давать надежные данные, если она достаточно стабильна.

Формирование содержательной части форм отчетности осуществляется с учетом требований Государственной программы перехода Российской Федерации на принятую в международной практике систему учета и статистики (1991—1996 гг.) и реализации федеральной целевой программы «Реформирование статистики в 1997—2000 годах».

Данные статистической отчетности поступают от предприятий и организаций в органы государственной статистики — либо в районные или городские отделы, либо прямо в областные (краевые) комитеты. После проверки данные разрабатываются в вычислительном центре (ВЦ): составляются сводные таблицы по формам, установленным Госкомстатом России. Данные обобщаются по отраслям, организационно-правовым формам, формам собственности, территориям и т.д. Сводные таблицы из местных статистических органов отправляются в Госкомстат России, где составляются сводные таблицы по стране в целом, рассчитываются сводные показатели с учетом тех же группировок данных (по отраслям, территориям, формам собственности и т.д.).

В настоящее время независимо от отрасли крупные и средние организации отчитываются по унифицированной отчетности.

Малые предприятия с 1999 г. отчитываются ежеквартально по форме № ПМ «Сведения об основных показателях деятельности малого предприятия». Унифицированная отчетность распространяется лишь на малые предприятия государственной формы собственности и собственности общественных организаций.

Остановимся на содержании унифицированных форм статистической отчетности исходя из редакции, утвержденной постановлением Госкомстата России № 67 от 17.07.2000 г.

Форма № П-1 «Сведения о производстве и отгрузке товаров и услуг» предусматривает отражение данных об объеме производства в целом по всем видам экономической деятельности. Наличие показателя общего объема производства создает основу для сопоставления и обобщения данных по предприятиям, занятым различными видами деятельности. Из общего объема производства выделяются производство товаров и производство услуг. При этом под *товарами* понимаются физические предметы, на которые могут быть распространены права собственности, а под *услугами* — проведенная по заказу деятельность, приводящая к изменению свойств или перемещению предметов, принадлежащих потребителю (например, перевозка, ремонт, хранение), либо к изменению состояния самого потребителя услуг (например, образовательные или медицинские услуги). К услугам относятся следующие виды дея-

тельности: оптовая и розничная торговля, ремонт зданий и сооружений, машин, оборудования, предметов личного пользования, транспорт и связь; услуги, связанные с недвижимым имуществом, арендой, исследовательской и коммерческой деятельностью, услуги в области образования и здравоохранения, коммунальные услуги и некоторые другие виды деятельности.

В форме № П-1 отражаются сведения о производстве и отгрузке конкретных видов товаров и услуг. В этих строках статистик должен правильно указать код товара (услуги) в соответствии с Общероссийским классификатором видов экономической деятельности (ОКВЭД) или коды Общероссийского классификатора продукции (ОКП) и Общероссийского классификатора услуг населению (ОКУН). По этим сведениям статистики относят приведенные данные к промышленности, платным услугам населению, розничной или оптовой продаже товаров.

Форма № П-2 «Сведения об инвестициях» содержит данные об инвестиционной деятельности предприятия (организации). Инвестиционная деятельность определяется как приобретение ресурсов, способных обеспечить получение доходов в будущем. В зависимости от типа приобретаемых активов инвестиции подразделяются на финансовые вложения, осуществляемые с целью приобретения финансовых прав (акций, облигаций и т.п.), и инвестиции в нефинансовые активы (здания, машины, землю и т.п.). Терминология, используемая в форме № П-2, приближена к определениям системы национальных счетов.

В формах № П-1 и П-2 все показатели даются в фактических ценах. Исключение из форм показателей в сопоставимых ценах не означает отказа от их использования. Пересчет в сопоставимые цены осуществляется не работниками предприятия, а в органах статистики по единой методологии, утвержденной Госкомстатом России, что обеспечивает большую точность и достоверность расчетов.

Введение унифицированных форм отчетности способствует автоматизации обработки и широкому использованию технологий работы с банками данных, основополагающими принципами которых являются упрощение показателей и снижение нагрузки на обследуемые предприятия.

В унифицированной форме № П-1 и в разделе формы № П-4 отсутствуют кумулятивные показатели «за период с начала года», ранее всегда присутствовавшие в формах статистической отчетности по продукции, численности и заработной плате работников. Это соответствует международной статистической практике, обычно оперирующей за отчетный и предшествующий ему месяцы, что облегчает заполнение форм в условиях значительной изменчивости экономических параметров. Кроме того, это позволяет исключить несоответствие между данными «за период с начала года» и суммой помесячных данных за соответствующее количество месяцев.

В новые формы включен ряд показателей, позволяющих выявлять наличие на предприятиях определенных экономических явлений. Например, заполнение строк 15 и 16 в форме № П-1 свидетельствует о том, что данное предприятие осуществляет экспорт или импорт услуг. Для более детального изучения структуры и направлений экспорта и импорта услуг такому предприятию высылается специализированная форма статистического наблюдения № 8-ВЭС (услуги) «Отчет об экспорте (импорте) услуг во внешнеэкономической деятельности». Аналогичную роль исполняют следующие показатели: вывоз товаров в государства — члены таможенного союза (строка 14 формы № П-1), инвестиции за рубеж и инвестиции из-за рубежа (строки 19, 20, 23, 24 формы № П-2).

Сопоставление показателей выпуска и отгрузки продукции в форме № П-1 позволяет сделать выводы от эффективности работы предприятия (продумана ли система реализации продукции без задержек или же предприятие работает «на склад»). При этом предусмотрены достаточно подробная классификация производимой продукции — строительно-монтажные работы, оборот торговли, потребительские товары и информация по каждому виду продукции в соответствии с перечнем, определенным органами государственной статистики. Некоторые данные относятся к будущему периоду и приводятся в форме № П-1 справочно. Например, информация об общем объеме заказов на поставку продукции в последующие периоды.

Форма № П-2 «Сведения об инвестициях» содержит показатели, которые позволяют анализировать масштабы и эффективность инвестиционной деятельности предприятия.

Раздельно отражается несколько групп инвестиций: в финансовые и нефинансовые активы; осуществляемые предприятием и в предприятие; краткосрочные и долгосрочные.

Направленность инвестиционной деятельности характеризуется как по отраслям вложений (промышленность, сельское хозяйство, строительство, транспорт, связь), так и по характеру вложений (паи, акции, облигации, займы). Раздельные данные по инвестициям в нефинансовые активы и финансовые инвестиции позволяют проследить процесс воспроизводства основных фондов.

Данные об источниках инвестиций позволяют определить долю собственных средств, вовлеченных в процесс инвестирования (это прежде всего прибыль предприятия), степень использования заемных средств (в том числе кредиты коммерческих банков, что является индикатором сбалансированности денежно-кредитного и реального секторов экономики).

Финансовое положение является интегральным показателем, характеризующим эффективность деятельности предприятия. Этим обусловлено выделение данных о финансово-хозяйственной деятельности *в отдельную форму № П-3 «Сведения о финансовом состоянии предприятия»*. В форме приведены данные о прибыли (убытке) предприятия за период с начала отчетного года, которые можно сопоставить с результатом за соответствующий период прошлого года (помещенного здесь же). На основе данных формы возможно сопоставление структуры дебиторской и кредиторской задолженности по срокам; приведена общая сумма оборотных средств (с выделением собственно денежных средств).

Таким образом, в форме предусмотрены сведения, требуемые для анализа финансовой устойчивости предприятия. Тут же отражаются данные о состоянии расчетов с предприятиями России, стран СНГ и других стран, что позволяет оценить степень вовлеченности предприятия в мировой рынок.

Эффективность деятельности предприятия определяется не только его производственными и финансовыми возможностями, значительное влияние оказывает его трудовой потенциал. *В форме № П-4 «Сведения о численности, заработной плате и движении работников»* численность работников приводится с разбивкой по категориям: работники списочного состава, внештатные совместители и выполнявшие работы по

Таблица 2.1

Сравнение запланированного и фактического круга единиц при контроле материалов наблюдения

Круг единиц			Итого
планируемый	фактический		
	соответствуют плановому кругу	вне планового круга	
Соответствуют фактическому кругу	n_{11}	n_{12}	$n_{1.}$
Вне фактического круга	n_{21}	n_{22}	$n_{2.}$
Итого	$n_{.1}$	$n_{.2}$	n

договорам гражданско-правового характера. Параллельно приводятся данные о форме начисленной заработной платы и выплатах социального характера по каждой из категорий.

Кроме того, приведены данные об эффективности использования времени (количество отработанных человекочасов и т.п.) и движении работников (принято работников, выбыло работников), а также данные о числе рабочих мест, намеченных к высвобождению.

Статистические наблюдения, основанные на системе унифицированных форм, позволяют получить комплексное отражение экономических процессов на предприятии — производственной, финансовой и инвестиционной деятельности.

Сличают список единиц, фактически охваченных обследованием, со списком, который был составлен до обследования. Несовпадения выявляются с помощью табл. 2.1.

Очевидно, что численность единиц, обозначенная n_{11}, составит основную часть объекта наблюдения. Численность n_{12} — те единицы, которые могли оперативно возникнуть до проведения наблюдения. Например, вновь созданные организации малого бизнеса, или мигранты, не учтенные паспортной службой. Конечно, эти единицы должны войти в объект наблюдения. Численность единиц n_{21} может относиться к тем единицам, которые перестали существовать на данной территории. Наконец, численность n_{22} должна быть равна нулю, т.е. таких единиц не было ни в плановом, ни в фактическом списке. Тем не менее эта численность, как показывает опыт,

не равна нулю и может составлять $(n_{22}/n) = 5–10\%$ всего массива данных.

2.6. Ошибки статистического наблюдения. Методы контроля данных наблюдения

Как бы тщательно ни был составлен инструментарий наблюдения, проведен инструктаж исполнителей, материалы наблюдения всегда нуждаются в контроле. Это объясняется массовым характером статистических работ и сложностью их содержания.

Прежде всего проверяется *полнота охвата единиц наблюдением*. С этой целью проводится сверка данных по спискам предприятий и организаций, ЕГРПО; пересчитываются заполненные анкеты. При проведении массовых социологических обследований часто, кроме основного списка, составляется дополнительный список респондентов на тот случай, если респонденты из основного списка почему-либо не могли быть опрошены. Дополнительный список формируется так, чтобы при замене респондентов общая структура опрашиваемых сохранялась. Поэтому при проверке устанавливается соответствие фактически опрошенных основному и дополнительному спискам. Проверка полноты охвата единиц не связана применением только сплошного наблюдения. В ходе проверки выявляются недоучет или повторный счет и обеспечение проектируемых пропорций собранных данных.

Одновременно на этой стадии проверяется *полнота заполнения каждого формуляра наблюдения* — формы отчетности, анкеты и т.д. После такой общей проверки проводится детальная проверка каждого полностью заполненного формуляра. При использовании нескольких формуляров проверка облегчается, если формуляры напечатаны на бланках разного цвета.

Для того чтобы хорошо организовать проверку, нужно представлять характер возможных ошибок. Все ошибки наблюдения можно назвать *ошибками регистрации*. Но они имеют разный характер и по-разному сказываются на результатах статистического исследования. Ошибки могут быть *случайными* и *систематическими*. Те и другие чаще всего возникают при опросе, но могут быть допущены и при непосредственном или документальном наблюдении.

Во всех случаях источником ошибок может быть как информация, поступившая от объекта наблюдения (ошибки в ответах опрашиваемого, искажения в показаниях приборов, регистрирующих какие-либо свойства объекта, ошибки в учетных документах), так и ошибки регистратора или экономиста предприятия, представляющего данные (неправильная запись ответов опрашиваемого, ошибки при переносе на формуляры наблюдения показателей приборов, данных учетных документов).

Случайные ошибки не имеют какой-либо направленности. Это описки, оговорки, перестановки цифр при записи цифровых данных и т.д. При обобщении массового материала они взаимопогашаются и не могут исказить значения сводных показателей и результаты анализа.

Другое дело *систематические ошибки* — они являются неслучайными и имеют определенную направленность. Такие ошибки очень опасны, поскольку приводят к искажению результатов статистического исследования. Эти ошибки, как правило, являются преднамеренными. Известно, например, что люди предпочитают преуменьшать свои доходы, округлять возраст, стараются показать большую осведомленность в области культуры, науки, чем есть на самом деле. Предприятия также могут внести элементы недостоверности в свою информацию, особенно в те характеристики, от которых зависят величина налоговых платежей, расчеты с кредиторами и т.п. Все ошибки такого рода необходимо выявить и исправить. Поэтому после проверки полноты данных проводится их контроль — счетный и логический.

Счетный контроль основан на жесткой связи между признаками, которая может быть проверена арифметическими действиями: сложением, вычитанием, умножением, делением. Связь такого рода часто отражается в заголовках граф отчетности и в подсказах: графа *X* равна графе *Y* плюс графа *Z* или графа *X* равна графе *Y*, деленной на графу *Z*, и т.д. Счетный контроль используется для проверки итоговых сумм. Если представленное число слагаемых не является полным, то сумма слагаемых должна быть меньше либо равна общему итогу, но не может превышать его.

Счетный контроль совершенно определенно устанавливает наличие ошибки, тогда как логический может лишь поста-

вить под сомнение правильность данных. *Логический контроль* основан на логической взаимосвязи между признаками. Классическим примером является взаимосвязь данных при переписи населения: вопросы о возрасте, образовании, семейном положении взаимоконтролируются. Если, например, окажется, что гражданин десяти лет женат, то ясно, что при заполнении формуляра допущены ошибки либо при записи возраста, либо другой характеристики.

Логический контроль основан и на сравнении с данными прошлого периода. Например, достоверность данных о выпуске продукции по видам может быть проведена сравнением с данными прошлого периода для того же предприятия. Кроме того, логический контроль опирается на представление о пределах возможных значений признака: минимуме и максимуме. Скажем, при проверке отчетности по форме № П-3 можно прикинуть, каким будет срок погашения дебиторской задолженности.

$$\text{Срок погашения дебиторской задолженности} = \frac{\text{Количество дней в отчетном периоде}}{\text{Оборачиваемость дебиторской задолженности}},$$

$$\text{Оборачиваемость дебиторской задолженности} = \frac{\text{Чистая выручка от реализации}}{\text{Средняя сумма дебиторской задолженности}}$$

Величина оборачиваемости дебиторской задолженности выражается в разах. Маловероятно, чтобы этот показатель был меньше 5 или больше 12 за год. При проверке срока погашения дебиторской задолженности мы можем использовать и нормативное значение этой величины (обычно 30 дней). Если реальный срок погашения намного (на несколько недель) отличается от нормативного в ту или иную сторону, необходимо поставить под сомнение резко отличающиеся данные и сделать запрос на предприятие.

Обычно для проверки поступающего материала наблюдения составляется *схема контроля*, в которую включаются все увязки между вопросами программы наблюдения — как арифметические, так и логические.

Никогда не следует произвольно вносить исправления в формуляр. Необходимо либо самому статистику провести повторное наблюдение (повторный опрос и т.д.), либо обратиться к лицам, отвечающим за предоставленную информацию (директору, главному бухгалтеру предприятия).

Данные наблюдения считаются принятыми, если они прошли контроль и если в них внесены исправления (по мере необходимости). Проверкой данных завершается начальный этап статистического исследования.

2.7. Реформирование российской государственной статистики

Национальная статистическая служба России совместно с другими федеральными органами решила следующие задачи:

- перешла на макроэкономические показатели, адекватные рыночной экономике;
- создала систему национальных счетов (СНС);
- разработала межотраслевые балансы по схеме СНС;
- организовала работы по международным сопоставлениям валового внутреннего продукта и разработала методику расчета валового регионального продукта;
- преобразовала финансовую, банковскую и бюджетную статистики;
- разработала платежные балансы;
- привела в соответствие с международными требованиями показатели статистики цен, статистики населения и труда, внешней торговли, включая таможенную статистику;
- участвовала в перестройке в соответствии с международными требованиями бухгалтерского и банковского учета;
- сформировала Единый государственный регистр предприятий и организаций всех форм собственности и хозяйствования;
- разработала и внедрила единую систему классификации и кодирования в соответствии с международными стандартами;
- осуществила подготовку и переподготовку кадров статистики и учета;
- развернула международное сотрудничество.

Все эти направления работ развивались в рамках «Государственной программы перехода Российской Федерации в 1992—1995 гг. на принятую в международной практике систему учета и статистики в соответствии с требованиями развития рыночной экономики». В 1995 г. в основном был завершен начальный этап реформирования российской стати-

стики. В ходе реализации Государственной программы была существенно изменена действующая система статистических показателей, создана система национальных счетов, приведены в соответствие с международной практикой показатели статистики финансов, населения, труда, внешней торговли, заложена основа Единого государственного статистического регистра юридических лиц и их обособленных подразделений, а также Единой системы классификации и кодирования технико-экономической и социальной информации.

В результате первого этапа реформирования национальной статистике удалось интегрироваться в мировое статистическое сообщество — мы научились говорить с цивилизованным миром на едином языке статистики.

Для продолжения реформ постановлением Правительства РФ № 1410 от 23 ноября 1996 г. была утверждена федеральная целевая программа «Реформирование статистики в 1997—2000 годах».

На втором этапе началось развитие функций ЕГРПО как инструмента, предназначенного для определения совокупности единиц статистического наблюдения, организации самого наблюдения и обработки отдельных статистических данных. Приказом Госкомстата России № 238 от 27 ноября 1996 г. «Об организации статистического наблюдения на основе ЕГРПО» в 1997 г. было предусмотрено проведение статистических наблюдений на основе генеральной совокупности (ГС). Формирование ГС-97 позволило создать основу для перехода к интегрированному принципу сбора отчетной информации и проведению статистического анализа по сопоставимому кругу объектов.

Начиная с 1998 г. все крупные и средние предприятия страны стали представлять в органы статистики отчеты на бланках унифицированных форм государственного статистического наблюдения № П-1, П-3 и П-4 (ежемесячно) и П-2 (ежеквартально). Внедрение в практику унифицированных форм статистической отчетности позволило обеспечить комплексный подход к анализу деятельности хозяйствующих субъектов независимо от отраслевой принадлежности и отказаться от большого числа узкоспециализированных форм отраслевой статистической отчетности.

Внедрение цензового принципа организации учета (крупные и средние предприятия) и переход от отраслевого метода

сбора информации к статистике предприятий представляют наиболее заметный результат реформирования государственной статистики во второй половине 1990-х годов.

Статистика предприятий предполагает формирование информационно-статистической базы, содержащей сведения о хозяйствующих субъектах всех отраслей экономики.

Предприятие представляет собой организационную единицу, производящую товары или услуги, пользующиеся определенной степенью автономии в принятии решений. В настоящее время предприятие рассматривается в России в качестве статистической единицы, а также в качестве правовой единицы как юридическое лицо. Четыре унифицированные формы отчетности предназначены, как уже отмечалось, для крупных и средних предприятий.

На базе отчетности органами государственной статистики осуществляются досчеты всех основных экономических показателей промышленности, строительства, сельского хозяйства, торговли, платных услуг населению, объемов перевезенных грузов и грузооборота автомобильного транспорта. Досчитываются до полного круга показатели деятельности малых предприятий и др.

Ежемесячно выполняются расчет занятого населения, досчеты численности работающих и средней заработной платы до полного круга, уровня общей безработицы, соотношения 10% высокооплачиваемых и 10% низкооплачиваемых работников и т.д.

Государственные статистические службы выполняют большой объем работ в форме единовременных специальных обследований. Примерами такого рода работ служат Всероссийская перепись населения 2002 г.; периодическое выборочное обследование населения по проблемам занятости; выборочное обследование инвестиционной активности предприятий в 2000 г.; сплошное единовременное обследование малых предприятий по результатам работы за 2000 г., проведенное в январе—апреле 2001 г. В результате последней работы была получена информация от 631,4 тыс. малых предприятий всех отраслей экономики, или от 76% числа предприятий, которые согласно сформированным перечням должны быть охвачены этой работой.

Обследование малых предприятий позволило получить развернутую информацию о параметрах малого предпринимательства в Российской Федерации, обобщающую характеристики развития малого бизнеса по видам экономической деятельности, административно-территориальным образованиям (городам, районам); определить место малого предпринимательства в экономике России. Расширяется область проведения экономических и сельскохозяйственных переписей как важнейшего инструмента получения информации в условиях становления рыночной экономики.

В период реформирования государственной статистики получила развитие ***гендерная статистика*** — статистика о женщинах и мужчинах, отражающая их положение во всех сферах жизни.

К концу 1998 г. в региональных комитетах по статистике закончилось формирование базы данных «Генеральная совокупность объектов статистического наблюдения» (БД ГС) на основе ЕГРПО. Госкомстатом России утверждено Положение о БД ГС, которое определило статус, структуру и источники информации, установило принципиальные подходы к ее формированию и ведению, порядок хранения, функции подразделения, ответственного за БД ГС, регламентировало представление данных из ГС. Ведется работа по созданию муниципальной статистики.

С переходом на унифицированные формы государственного статического наблюдения в органах статистики активизировался переход к новому поколению ЭВМ: от первых отечественных ПЭВМ типа ЕС-1840, появившихся в 1988 г., IBM PC IT 286 — в 1991 г., до Pentium — в 1999 г. Начался качественно новый этап в развитии информационных технологий. Был внедрен проект «Технология-2000». В результате многие комплексы электронной обработки статистической информации были переведены на ПЭВМ, а загрузка больших ЭВМ в течение 1997—1998 гг. сократилась практически до нуля.

Важнейшим событием стало внедрение средств электронной почты для связи региональных комитетов с Госкомстатом России — электронная почта превратилась в основной канал обмена информацией. Были освоены программные продукты ЭП «1С», ГАС «Выборы», ЭП Госналогслужбы. Начиная с 1998 г. используется Интернет: адрес сайта Госкомстата России: **www.gks.ru**.

Работник государственной статистики имеет статус государственного служащего.

РЕЗЮМЕ

Получение статистических показателей, характеризующих экономику, население страны и отдельных регионов, составляет обязанность государственной статистической службы — так называемой официальной статистики. Деятельность учреждений государственной статистики финансируется из бюджета РФ.

Система государственной статистики России сложилась во второй половине XIX в. Статистическая деятельность организована по принципу централизации. Во главе государственной статистики стоит Государственный комитет по статистике Российской Федерации (Госкомстат России). Статистическая работа ведется также министерствами, государственными комитетами. Это ведомственная статистика.

Региональная статистика формируется из системы статистических показателей, рассчитываемых для всех регионов, в соответствии с Федеральной программой статистических работ, а также из статистических показателей, отражающих региональные особенности, получение и разработка которых финансируются из местного бюджета.

Одна из важнейших функций государственной статистики — регулярная публикация статистических показателей в виде статистических сборников.

Государственные статистические службы связаны с международными статистическими организациями, обеспечивающими сопоставимость данных по разным странам, регулярность выполнения переписей и обследований, разработку и внедрение методологии статистических работ. В области статистической практики выделяются следующие основные международные организации: Статистическая комиссия ООН, статистические отделы региональных экономических комиссий, Евростат.

В области статистической науки, образования, компьютерной поддержки статистики, развития и методологии официальной статистики ведущую роль играет Международный статистический институт, созданный в 1885 г.

Статистическое наблюдение — научно организованный сбор данных по единицам (или группам единиц) совокупности с целью их последующего обобщения.

Собираемые данные должны быть достоверными и сопоставимыми. Первое требование определяет объективность статистики, второе — возможность агрегирования данных отдельных единиц в сводные статистические показатели.

Статистическое наблюдение подразделяется на виды по времени и по охвату единиц наблюдения.

Программа статистического наблюдения зависит от поставленной цели.

Объект наблюдения определяется с позиций единиц наблюдения, территории и времени наблюдения.

Единица наблюдения — явление, признаки которого подлежат регистрации.

Инструментарий наблюдения включает формуляр(ы) наблюдения и инструкцию по заполнению формуляров наблюдения. Формуляр наблюдения — документ единого образца, содержащий программу наблюдения и предусматривающий занесение соответствующих ей данных.

Перепись — главный вид специального наблюдения. Как правило, перепись — это сплошное статистическое наблюдение. Критический момент переписи — день года и час, по состоянию на который должны быть определены границы объекта наблюдения и проведена регистрация признаков по каждой единице наблюдения. В нашей стране расширяется область экономических и сельскохозяйственных переписей.

Период наблюдения — время, в течение которого проводится заполнение формуляров наблюдения.

По крупным и средним предприятиям статистическое наблюдение проводится в форме унифицированной отчетности, не зависящей от специфики отрасли.

Материалы статистического наблюдения подвергаются контролю с точки зрения полноты охвата единиц, полноты регистрации признаков, достоверности данных, которая проверяется методами логического и счетного контроля.

Качество материалов наблюдения зависит от частоты и характера ошибок. Ошибки наблюдения подразделяются на случайные (ошибки регистрации) и систематические. Первые не сказываются на значениях сводных показателей, так как

взаимно погашаются при обобщении данных; вторые приводят к искажению сводных показателей.

Структура, направление и содержание работ отечественной государственной статистики существенно реформированы с 1992 г. в целях внедрения международных стандартов, освоения методологии построения системы национальных счетов, расчета макроэкономических показателей, принятых в мировом сообществе. Переход на международные стандарты в области учета и статистики составляет необходимое условие вхождения России в международные экономические организации, участия в международных проектах, получения права заимствования на определенных условиях.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Воронов Ю. П.* Методы сбора информации в социологическом исследовании. — М.: Статистика, 1974.
2. *Деев Г., Крутова Т.* Метод основного массива в статистических наблюдениях // Вестник статистики. — 1992. — №5. — С. 39—43.
3. *Деев Г., Мухин П.* Несплошное статистическое наблюдение: исторический опыт, практика, перспективы // Вопросы статистики. — 1996. — №3. — С. 21—27.
4. *Елисеева И. И.* Моя профессия — статистик. — М.: Финансы и статистика, 1992.
5. *Елисеева И. И., Костеева Т. В., Хоменко Л. Н.* Международная статистика. — Минск: Вышэйшая школа, 1995.
6. *Курс* социально-экономической статистики /Под ред. М. Г. Назарова. — М.: Финстатинформ, 2002.
7. *Моргенштерн О.* О точности экономико-статистических наблюдений. — М.: Статистика, 1968.
8. *Об ответственности* за нарушение порядка представления государственной статистической отчетности // Вестник статистики. — 1992. — №1. — С. 3—7.
9. *Основные* итоги работы Госкомстата России по реформированию государственной статистики в 1997—2001 годах. — М.: Госкомстат России, 2002.
10. *Рябушкин Т.В., Симчера В.М.* Очерки международной статистики. — М.: Наука, 1981.
11. *Суринов А. Е.* Официальная статистика в России: проблемы реформирования. — Tacis. РЕЦЕП, 2002.

Глава

3

СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ

3.1. Сущность и значение статистических показателей. Показатель и его атрибуты

В главе 1 сказано, что статистика выражает массовые явления и процессы количественно в числовой форме. Но «числа», применяемые в статистике, это не абстрактные числа математики, характеризуемые только величиной, знаком, формой (целые — дробные; мнимые — действительные; рациональные — иррациональные и т.п.). Статистика применяет, собственно говоря, не числа, а показатели, точнее статистические показатели.

Что же такое статистический показатель? каковы его содержание и построение? какие виды показателей используются в статистике? какое значение имеют статистические показатели в познании массовых явлений и процессов, в управлении производством, в жизни общества в целом? Ответам на эти вопросы посвящена данная глава.

Не умея правильно понять содержание, форму, свойства того или иного статистического показателя, нельзя корректно применить его в анализе социально-экономических явлений и процессов и осознать смысл статистической информации и жизни страны и мира.

С философской точки зрения статистический показатель — это *мера*, т.е. единство качественного и количественного отражения свойств объективных явлений и процессов в научном сознании. Поскольку статистика изучает массовые явления, *статистический показатель* — это *обобщающая характеристика какого-то свойства совокупности, группы.* Этим он отличается от индивидуальных значений, которые, как отмечалось, называются признаками. Например, средняя продолжительность ожидаемой жизни родившегося поколения людей в стране — статистический показатель. Продолжительность жизни конкретного человека — признак.

Рассмотрим содержание и форму статистического показателя на примере ввода в действие жилых домов в Российской Федерации в 2000 г., составившего 30,3 млн $м^2$ общей площади. Показателем является не только число 30,3, а весь текст, поясняющий его содержание. Качественная сторона этого показателя — ввод в действие жилых домов. Статистический показатель имеет и количественную сторону, которая выражается числом и единицей измерения: 30,3 млн $м^2$ общей площади.

Не всегда статистический показатель является именованным числом. Он может быть абстрактным и отвлеченным числом без наименования, может быть выражен в долях единицы: в процентах, промилле и т.п. Именованными числами являются абсолютные статистические показатели.

Статистический показатель имеет указание на территориальные границы объекта (жилье на определенной территории — Российской Федерации) и границы во времени — 2000 г. Без указания территориальных, отраслевых или ведомственных границ объекта и без привязки к определенному интервалу времени или моменту статистический показатель не существует. Атрибуты статистического показателя представлены на рис. 3.1.

Являясь отображением свойств изучаемых явлений и процессов, статистический показатель служит орудием их познания. Но всякое знание всегда ограничено, неполно соответствует изучаемому объекту. Ни один статистический показатель, ни целая их *система* не могут отразить все свойства, все

Качественная сторона: объект, его свойство	Количественная сторона: число и единицы измерения	Территориальные, отраслевые и иные границы объекта	Интервал или момент времени

Рис. 3.1. Атрибуты статистического показателя

особенности объекта и даже часть этих свойств с абсолютной точностью. Статистический показатель — приближенное, неточное и неполное отображение свойств изучаемого объекта, доступное при имеющемся уровне знаний и возможностях учета, измерения, сбора и передачи информации. Каждому ясно, что невозможно точно измерить вес собранного картофеля без примеси песка, глины, частиц почвы и камней, невозможно в масштабах целой республики избежать ошибок во взвешивании, записи, передаче сведений об урожае. Это один из наглядных примеров. Известно, что бывают сознательные искажения данных — приписки. Если же речь идет о жизни общества, как-то: уровень материального благосостояния, эффективность производственного процесса, культурный уровень населения, то главной причиной неточности, неполноты отображения этих сторон общественной жизни статистическими показателями являются недостаточное развитие тех наук, которые формируют указанные категории, и трудности перехода от их качественного описания к количественному измерению.

Поэтому статистические показатели не есть нечто раз навсегда застывшее. Одни развиваются, улучшаются, от иных отказываются за ненадобностью, создаются новые. Так, в настоящее время мир наводнен *рейтингами*, которые также могут рассматриваться как обобщающие показатели.

Признак и показатель

Остановимся на соотношении между признаком и статистическим показателем.

Признак — это свойство, присущее единице совокупности. Признак входит в качественное содержание показателя, он существует объективно независимо от того, отражает ли его наука с помощью тех или иных показателей. Например, возраст человека — это его признак, который можно измерять с разной степенью точности — в годах, месяцах, в сутках или охарактеризовать датой рождения.

Показатель — характеристика группы единиц, или совокупности в целом. Его построение зависит от цели исследования и изобретательности статистика. Средний возраст работников фирмы или жителей города — это статистические

показатели, дающие возрастную характеристику определенных групп. Другим видом возрастных показателей могут служить ряд распределения людей по возрасту и вычисленные на основе этого ряда системы показателей для характеристики структуры такого ряда и размеров вариации (гл. 5).

3.2. Классификация статистических показателей

Объектами статистического исследования могут быть самые разнообразные явления и процессы. Поэтому чрезвычайно велико и разнообразие статистических показателей. В данном разделе рассматривается только наиболее общая классификация статистических показателей (табл. 3.1). Их конкретные виды и формы представлены в последующих главах учебника, в курсах математической, социально-экономической и отраслевых статистических дисциплин.

Таблица 3.1

Классификация статистических показателей

По качественной стороне показателей	По количественной стороне показателей	По отношению к характеризуемому свойству
Показатели свойств конкретных объектов Показатели статистических свойств любых массовых явлений и процессов	Абсолютные Относительные	Прямые Обратные

Показатели конкретных свойств изучаемого объекта — это, например, уже упомянутый средний возраст работников предприятия, объем реализованной продукции предприятия, валовой внутренний продукт государства, средний надой молока на корову на ферме, объем перевозок груза автопарком, показатели рождаемости, смертности, обеспеченности населения товарами и услугами, национальное богатство, средний душевой доход жителя страны и т.д. Особенностью этих показателей является то, что они формируются не только статистикой. В построении этих показателей их качественное содержание определяется конкретной предметной наукой: показатель рождаемости — демографией, показатель внутреннего валового продукта — теорией экономики, показатели уро-

жайности, продуктивности скота — соответствующими сельскохозяйственными науками. Статистика отвечает за методику учета или расчета количественной стороны этих показателей и их форму.

Совершенно иначе обстоит дело с *показателями статистических свойств* любых массовых явлений и процессов, не зависящих от конкретного содержания этих явлений. К таким статистическим показателям относятся: средние величины, показатели вариации, показатели связи признаков, показатели структуры и характера распределения, показатели скорости и темпов изменения, показатели колеблемости в динамике; статистические оценки степени точности и надежности любых конкретных статистических показателей, полученных при выборочном изучении совокупности, а также оценки надежности и точности статистических прогнозов. За качественную и количественную сторону этих показателей, за их построение, интерпретацию и применение отвечает не какая-либо иная научная дисциплина, а только сама статистика. Это, можно сказать, ее кровные дети! Система таких показателей создается и совершенствуется в ходе развития методов статистики, поэтому в последующих главах будут рассмотрены построение, свойства и применение именно таких статистических показателей.

Теоретическая статистика разрабатывает и изучает содержание, форму, методы расчета этих показателей в общем виде: что такое средняя арифметическая величина, коэффициент вариации, уравнение тренда ряда динамики. Если же любой из этих показателей рассчитан для определенного объекта, признака, периода времени, то он становится уже конкретным показателем. Статистические показатели подразделяются на абсолютные и относительные.

Абсолютным показателем является такой, который отражает либо суммарное число единиц, либо суммарное свойство объекта. Например, число фермерских хозяйств в Ленинградской области на 1 января 2003 г., посевная площадь картофеля в районе, сумма средств, направленных на потребление за конкретный месяц или год, и т.п.

Абсолютные показатели, как правило, выражаются именованными величинами в натуральных единицах измерения: тоннах, штуках, часах, амперах и т.п., в условных единицах:

условном топливе, нормо-сменах, килономерах пряжи и т.д. или в стоимостных единицах: рублях, долларах, марках. Они характеризуют сумму значений первичных признаков объекта. Совершенно понятно, что наука не может ограничиваться характеристиками только изолированных свойств объекта. Поэтому статистика не ограничивается абсолютными показателями. Она измеряет и характеризует соотношение разных абсолютных величин, их изменения во времени, их взаимосвязи между собой и связи с окружающей средой. Статистика, как и все науки, широко пользуется общенаучными методами сравнения, обобщения, синтеза.

Относительным показателем является показатель, полученный путем сравнения, сопоставления абсолютных или относительных показателей в пространстве (между объектами), во времени (по одному и тому же объекту) или сравнения показателей разных свойств изучаемого объекта.

Относительные статистические показатели, получаемые при сопоставлении абсолютных показателей, могут быть названы относительными показателями *первого порядка*, а полученные при сопоставлении относительных же показателей — *показателями высших* (второго, третьего и т.д.) *порядков*. Показатели выше четвертого порядка ввиду сложности интерпретации почти никогда не применяются. Относительные статистические показатели выражают связь между абсолютными показателями: урожайность картофеля — отношение валового сбора к посевной площади; доля городского населения в стране — отношение численности населения городов к общему числу жителей страны.

Основные виды относительных величин чаще выражаются отвлеченными числами, но могут быть также именованными относительными показателями. Их построение связано с применением различных методов статистики.

Относительные показатели можно подразделить на следующие группы.

1. Относительные показатели, характеризующие структуру объекта. Это доля (удельный вес) — отношение части к целому. Например, отношение площади каждой из сельскохозяйственных культур к общей посевной площади; числа женщин к общей численности населения города, республики. В эту же группу входят характеристики отношения между отдельными

частями объекта; показатели, характеризующие степень сложности структуры, степень неравномерности (вариации) долей и др. Доли выражаются нередко в процентах или промилле (тысячных долях).

2. Относительные показатели, характеризующие динамику процесса, изменение во времени. Это отношения показателей, характеризующих объект в более позднее время (текущий период), к аналогичным показателям того же объекта в более ранний (базисный) период. Такие показатели называют *темпами роста*. Темп роста может быть выражен в разах или в процентах. Темп роста говорит о том, во сколько раз больше показатель текущего периода в сравнении с базисным или сколько процентов он составляет по отношению к показателю базисного периода. К относительным показателям динамики принадлежат также темпы прироста, параметры уравнений трендов, коэффициенты колеблемости и устойчивости в динамике, индексные показатели динамики. Подробнее о них сказано в главе 12.

3. Относительные показатели, характеризующие взаимосвязь признаков в совокупности явлений, а также взаимосвязь результативных признаков-следствий с факторными признаками-причинами. Например, связь уровня душевого дохода с размером потребления мяса или фруктов на одного человека; связь дозы удобрений с урожайностью картофеля и т.п. К таким показателям относятся рассматриваемые в главе 9 коэффициенты корреляции, эластичности, детерминации, а также аналитические индексы. Относительные показатели взаимосвязи могут быть как отвлеченными, так и именованными числами.

4. Относительные показатели, характеризующие соотношение разных признаков того же объекта между собой (иногда их называют показателями интенсивности). Эти показатели обобщают вторичные признаки объектов (например, производительность труда — отношение произведенной продукции в натуральном или стоимостном выражении к затратам труда на ее производство и др.). Показатели соотношения признаков могут быть прямыми и обратными. Например, отношение затрат труда на производство к объему продукции дает показатель трудоемкости продукции — величину, обратную прямому показателю производительности труда. И прямые, и обратные показатели выражаются именованными числами с двойными единицами измерения обоих сравниваемых признаков: в рублях за 1 час труда, в центнерах с 1 га площади. Например, продукция предприятия учитывается в миллионах рублей за год, скажем 1800, и стоимость основных производственных фондов предприятия тоже учитывается в миллионах рублей, скажем 4000. Если формально единицы измерения сравниваемых признаков совпадают, то неверно называть фондоотдачу — показатель сравнения стоимости продукции за год со стоимостью среднегодовых производственных фондов отвлеченным числом (в нашем примере — 0,45). Правильно будет сказать: «Фондоотдача составила 45 коп. продукции на 1 руб. основных фондов за год». Стоимость продукции и стоимость фондов — разные признаки, хотя имеют одинаковую единицу измерения.

В экономике относительные показатели, характеризующие величину признака объекта, рассчитанные на единицу другого признака, используются для измерения эффективности либо интенсивности производства.

К данному классу показателей принадлежат и показатели, характеризующие степень системности признаков, например соотношение между суммой осадков и суммой эффективных температур (способствующих произрастанию сельскохозяйственных культур), так называемый гидротермический коэффициент; таково же соотношение между весом и ростом человека, характеризующее пропорциональность его тела.

5. Особым видом относительных статистических показателей являются отношения фактически наблюдаемых величин признака к его нормативным, плановым, оптимальным или максимально возможным, величинам. Это широко распространенные на производстве показатели выполнения норм выработки, норм расхода материалов и других ресурсов. Отношения наблюдаемых величин признака к оптимальным или плановым характеризуют приближение изучаемого процесса к идеалу. Так, если оптимальная норма потребления мяса взрослым мужчиной на Северо-Западе России составляет 80 кг в год, а фактическое среднедушевое потребление составило в 1997 г. 54 кг, то ясно, что размер и структура потребления далеки от оптимального: всего 68%. Всякое превышение или недобор до оптимальной величины, всякое откло-

нение от 100% такого относительного показателя (в любую сторону) означают нарушение оптимальности процесса, даже перевыполнение плана, если план не лозунг, а научно обоснованная, взаимосвязанная система объемов производства отдельных видов продукции. В этом случае превышение планового выпуска одного вида продукции, например выплавки стали, без согласованного изменения производства станков, прокатных станов, других средств обработки металла есть попросту омертвление затрат и бесполезный перерасход природных ресурсов, труда.

Отношение фактических значений признака к максимально возможным значениям часто характеризует качество процесса, агрегата, машины. Таковы, например, коэффициенты полезного действия двигателей, электромоторов. Отношения фактических показателей вариации к максимально возможным при данной численности совокупности используются при анализе вариации (гл. 5), при измерении степени специализации предприятия или региона на производстве определенной продукции и в ряде других задач.

Само задание в той или иной отрасли экономики может быть выражено относительной величиной динамики или структуры. Например, «снизить затраты топлива на 1 кВт · ч электроэнергии на 5% в сравнении с прошлым годом»; «увеличить долю продукции высшего качества до 85% общего выпуска». Показатели выполнения такого задания будут являться относительными показателями второго порядка.

6. Еще один вид относительных статистических показателей возникает в результате сравнения разных объектов по одинаковым признакам. Сравнение урожайности одной и той же культуры в том же году между хозяйствами, областями; сравнение показателей производства или уровня жизни населения в разных странах — это обычные приемы познания. При построении таких относительных показателей необходимо позаботиться, чтобы сравниваемые показатели определялись по единой методике построения, были сравнимы по единицам измерения и во всех других отношениях. В социально-экономической статистике есть специальный раздел о международных сравнениях показателей.

Пример. Сравним производство валового внутреннего продукта на душу населения в Великобритании и в США в

1996 г.: в Великобритании на одного жителя было произведено 19 528 долл. (по паритету покупательной способности), в США — 27831 долл./чел. Показатель сравнения может быть построен как отношение одного душевого уровня к другому: душевое производство ВВП в Великобритании составило 70,2% душевого производства ВВП в США. Или душевое производство ВВП в США составило 1,425, или 142,5% душевого производства в Великобритании. Если речь идет об исследовании по экономике Великобритании, то предпочтительнее первая форма показателя: изучаемый объект (сравниваемая величина) — числитель, а другой объект (база сравнения) — знаменатель относительного показателя. Если изучается экономика США, предпочтительнее взять в числителе показатель США.

3.3. Общие принципы построения относительных статистических показателей

Построение относительных показателей — задача, требующая сочетания конкретного знания свойств объекта и общих закономерностей статистической методологии. Остановимся на общих логико-статистических принципах построения относительных показателей.

Первый принцип. Относительный показатель как сравнение двух абсолютных величин, которые объективно связаны, должен быть независим от нашего желания. Если этого условия нет, получится согласно русской поговорке «В огороде — бузина, а в Киеве — дядька». Связать этого «дядьку» с «бузиной» чисто математически, разделив одно число на другое, можно, но никакого относительного показателя мы не построим.

Необходимо добиваться как можно большего соответствия по смыслу сравниваемых показателей. Например, мы хотим построить относительный показатель, характеризующий степень грамотности населения. Можно разделить число грамотных на общую численность населения, но это не лучший из показателей. Ведь ясно, что дети до 6 лет, некоторые категории инвалидов с детства, душевнобольных не могут наравне со здоровыми и достигшими школьного возраста людьми быть обучены грамоте. Эти категории лиц правильнее исклю-

чить из всего населения при построении относительного показателя грамотности.

Пример. Продуктивность молочного скота определяется делением полученного валового надоя молока на маточное поголовье (коров, коз, овец); продуктивность в производстве яиц — делением валового сбора на поголовье кур-несушек (или уток, гусынь), не включая, разумеется, самцов-петухов, селезней, гусаков. Но если продуктивность в производстве шерсти мы станем определять путем отношения валового настрига шерсти на поголовье только овцематок и козоматок, то сделаем ошибку: ведь шерсть стригут и с баранов!

Второй принцип. При построении относительного статистического показателя сравниваемые величины могут различаться только одним атрибутом: или видом признака (при одинаковом объекте, периоде времени, плановом или фактическом характере показателей), или временем (при том же признаке, объекте и т.п.), или только фактическим, плановым или нормативным характером показателей (тот же объект, признак, время) и т.д. Нельзя сопоставлять показатели, различные по двум и более атрибутам, скажем, сравнивать добычу угля в США в 2000 г. с выплавкой стали в Российской Федерации в 2002 г.

Третий принцип. Необходимо знать возможные границы существования относительного показателя. Например, относительные показатели вариации теряют смысл и не могут применяться в тех случаях, когда их знаменатели — средние значения признаков — близки к нулю, потому что при стремлении знаменателя к нулю относительный показатель стремится к абсурдному бесконечному значению. Аналогично если исходные показатели в текущем и базисном периодах имеют разные знаки, то теряет смысл и не может применяться относительная величина динамики — темп роста. Если предприятие имело в 2000 г. убыток 150 млн руб., а в 2001 г. получило прибыль 300 млн руб., неверно ни то, что «финансовый результат вырос вдвое» (если отбросить знаки), ни то, что он «вырос в минус два раза», если делить +300 млн на −150 млн.

Относительные показатели, измеряющие степень приближения некоторого признака к предельному значению, должны строиться так, чтобы в пределе увеличения они стремились к единице, а в другом пределе своего уменьшения — к

нулю. Так строятся коэффициенты, измеряющие тесноту признаков, степень эффективности использования ресурсов, скажем, КПД двигателя. Для многих характеристик экономической, тем более социальной и экологической эффективности производственных процессов такие относительные показатели эффективности еще предстоит построить.

Ввиду того, что анализ структурных сдвигов в наше время имеет большое значение в экономике, относительные и абсолютные характеристики структуры и ее изменений подробно рассматриваются в данном учебнике (гл. 14).

Особая методика построения показателей необходима в тех случаях, когда сравниваемые показатели имеют разные знаки (прибыль «+», убыток «−») или один из них имеет нулевое значение.

3.4. Понятие о системах статистических показателей

Свойства, признаки изучаемых статистических объектов (совокупностей процессов) не изолированы, а связаны между собой. Поэтому и показатели этих свойств образуют более или менее полную систему. Число взаимосвязанных показателей может составлять от двух-трех до нескольких сотен.

Различают жестко детерминированные и статистические связи показателей. Примером системы жестко связанных показателей может служить система объемных и качественных показателей промышленности России за 1998 г.

Абсолютные показатели

1. Стоимость основных производственных фондов — 3982 млрд руб.
2. Численность промышленно-производственного персонала — 13 300 тыс. чел.
3. Объем продукции промышленности — 1681 млрд руб. за год.

Относительные показатели

1. Фондовооруженность персонала: $\frac{3982 \text{ млрд руб}}{13\,300 \text{ тыс. чел.}} =$ = 299,4 тыс. руб./чел.

2. Фондоотдача: $\frac{1681 \text{ млрд руб}}{3982 \text{ млрд руб}} = 0{,}422$ руб. продукции на 1 руб. фондов в год.

3. Производительность труда: $\frac{1681 \text{ млрд руб}}{13\,300 \text{ тыс. чел.}} =$ $= 126{,}4$ тыс. руб./чел. в год.

Каждый показатель этой системы может быть точно вычислен по остальным показателям, так как он является либо частным от деления других показателей, либо произведением показателей. Это означает, что жестко детерминированная система показателей может быть подвергнута арифметической проверке. Например, производительность труда должна быть равна произведению показателей фондовооруженности персонала и фондоотдачи:

299,4 тыс. руб./чел. · 0,422 в год = = 126,4 тыс. руб./чел. в год.

Объем продукции промышленности равен произведению трех показателей: численности персонала, его фондовооруженности и фондоотдачи:

13 300 тыс. чел. · 299,4 тыс. руб./чел. · 0,422 руб. продукции на 1 руб. фондов в год = 1681 млрд руб. в год.

Примером системы показателей, связанных статистической зависимостью (гл. 9), служит система факторов, влияющих на величину заработной платы рабочего:

- результативный показатель — средняя месячная заработная плата (руб./чел.);
- факторные показатели:
  - возраст рабочего;
  - стаж работы по данной специальности;
  - число отработанных часов в месяц;
  - выработка — число деталей или операций за час работы;
  - разряд или класс рабочего;
  - показатели рентабельности предприятия;
  - отрасль промышленности.

Никакие арифметические действия над величинами факторных показателей не приводят к величине результативного показателя. Его величина не может быть проверена арифметически. Однако средняя величина заработной платы в совокупности рабочих связана со стажем, с разрядом рабочего. Стаж, в свою очередь, связан с возрастом, рентабельность предприятия — с отраслью. Все показатели образуют систему, но связь их проявляется в среднем для достаточно большой совокупности рабочих.

Система статистических показателей, как правило, должна включать как абсолютные показатели, так и относительные. Изолированный абсолютный показатель подобен человеку в пустыне: он не говорит ничего, ибо ему не с кем говорить. Предположим, предприятие произвело продукции в 1999 г. на 46 млрд руб. Из этого показателя нельзя сделать никакого вывода, пока его величина не сопоставлена с числом работников, затратами на производство, объемом продукции за предыдущий год и т.п., т.е. пока этот показатель не будет включен в систему и не будут построены относительные величины. Из этого не следует делать заключение о большей информативности относительных показателей. Если известно, что в студенческой группе число отличников в данную сессию составило 200% к их числу в прошлую сессию, то это не значит, что группа резко повысила уровень знаний. Может быть, в прошлую сессию был 1 отличник из 27 человек, а теперь стало 2, что и составило 200%. Только сочетание абсолютных и относительных показателей позволяет достаточно полно характеризовать объект в отношении поставленной задачи его изучения.

3.5. Функции статистических показателей

О роли и значении статистики в развитии общества, в научном познании окружающего мира и в управлении предприятием, учреждением уже сказано в предыдущих главах учебника. Конкретизируем теперь эти вопросы применительно к системам и видам статистических показателей.

Основной функцией статистических показателей и их систем является *познавательная информационная функция.* Без статистической информации невозможны познание закономерностей природных и социальных массовых явлений, их предвидение, а значит, регулирование либо прямое управление, будь то на уровне отдельного предприятия, фермы, го-

рода или региона, на государственном или межгосударственном уровне. Отдельный человек или семья, не представляющие, сколько в среднем за месяц или за год они тратят на покупку продуктов питания, на обувь и одежду, на оплату коммунальных услуг, не могут рационально расходовать средства, планировать свой бюджет. Фермеру необходимо знать показатели средней урожайности за ряд лет различных сельскохозяйственных культур на его участках земли, показатели колеблемости и устойчивости урожаев в зависимости от условий погоды, среднюю частоту поломок деталей машин, средние цены (и темпы их роста) на покупаемые удобрения и т.д. Тем более попытки управлять государством субъективно, не опираясь на систему достаточно надежных статистических показателей — путь к социальной, экономической и экологической катастрофе.

Среди познавательно-информационных функций статистических показателей выделяется функция *мониторинга* — постоянно действующего наблюдения при постоянстве рассчитываемых показателей. Например, существует мониторинг Центрального банка России за деятельностью коммерческих банков или экологический мониторинг и т.д. Кроме того, показатели выполняют роль *системы сигналов*, свидетельствуя о социальной напряженности (число забастовок, в том числе по причинам, количество бастующих, процент неявок на выборы, уровень преступности и т.д.). В этой своей функции показатели отражают *экономическую безопасность страны, равномерность распределения* инновационных центров по территории страны и т.д. Условием выполнения статистическими показателями их информационной, познавательной функции являются их научное обоснование и достаточно точное и надежное, а также своевременное количественное определение.

Прогностическая функция, т.е. роль статистических показателей в предвидении будущего, тесно связана с их информационной функцией. Конечно, данная функция присуща не всем статистическим показателям, а тем из них, которые используются при моделировании массовых процессов.

Оценочная функция статистических показателей заключается в том, что на их основе люди, общество, государство оценивают деятельность предприятий, организаций, трудовых и

творческих коллективов, правительств. Великий немецкий писатель, поэт и мыслитель И. В. Гете за два года до смерти в разговоре со своим секретарем И. П. Эккерманом сказал: «Считают, будто числа управляют миром. Но я знаю, что числа учат нас узнавать, хорошо ли мир управляется»[1]. А российский статистик, автор учебника статистики в России К. Ф. Герман (1767—1838) писал: «Статистик есть публичный провозвестник и доброго, и худого, и контролер правительства»[2]. Да, по надежным «истинным» статистическим показателям, а не по речам и рекламным роликам население должно и может оценивать деятельность руководителей всех рангов. Но при этом недопустимо такую оценку давать по отдельному показателю, произвольно вырванному из системы. Долгое время в СССР деятельность предприятий оценивалась на основе показателя выполнения плана по валовой продукции. Поскольку в этот показатель включается и стоимость незавершенных изделий, то ради получения высокого показателя выполнения плана и премии к концу отчетного периода на предприятии авральо собирали шасси, не имея моторов, закладывали новые стройки, не достроив предыдущие, и т.д. Омертвление огромных материальных средств и труда — вот результат превращения отдельного статистического показателя в главное и единственное мерило успехов производства. Также неверно оценивать успешность развития экономики страны только по показателю низкой инфляции или только по внешнеторговому сальдо — по любому отдельно взятому статистическому показателю.

Рекламно-пропагандистская функция статистических показателей — еще более щекотливый вопрос. С одной стороны, реклама — это неотъемлемый атрибут рыночной экономики, и фирмы, компании, естественно, стремятся использовать в рекламе статистические показатели о долговечности, качественности своей продукции, зная, что цифровым данным люди доверяют больше, чем словам. Однако при таком использовании статистических показателей велик риск либо подмены реального показателя планируемым, т.е. желаемым, но еще не осуществленным, либо умолчания о других показателях товара, не отвечающих целям рекламы. Поэтому к стати-

[1] *Eckermann I. P.* Gespräche mit Goethe. — Leipzig, 1902. — S. 313.

[2] *Герман К. Ф.* Всеобщая теория статистики. — СПб., 1809. — П. 78.

стическим показателям, применяемым в рекламных целях, следует относиться весьма осторожно, по возможности проводить дополнительные расчеты и анализ. Например, фирма «Кудесник», рекламируя в газете «Известия» от 14 января 1997 г. кран КС-5579 на базе грузовика «КамАЗ», сообщила, что средний ресурс крана до капитального ремонта составляет 10 лет эксплуатации, или 8000 часов. Оба показателя впечатляют. Но если провести расчет, на какие же условия эксплуатации рассчитан этот ресурс, то выяснится, что на 1 год приходится 800 часов работы, на 1 месяц при 22 рабочих днях — 66 часов, на сутки — 3 часа работы. Неудивительно, что при столь низком показателе использования по времени — всего 0,375 одной смены в сутки кран, возможно, и проработает 10 лет без капитального ремонта.

Также осторожно следует подходить и к статистическим показателям, используемым государствами, политическими партиями, кандидатами на выборные должности в их агитации и пропаганде. Статистическая наука всегда честно указывает на ограничения, приближенность, вероятностный характер многих своих показателей, лишь постепенно, ограниченно приближающих нас к познанию бесконечно сложного окружающего мира.

РЕЗЮМЕ

Статистический показатель — это обобщающая характеристика какого-либо свойства совокупности, группы явлений.

Атрибуты статистического показателя включают определение качественной стороны характеризуемого свойства, количественное выделение этого свойства (числовая величина и единица измерения), территориальные, отраслевые и иные границы объекта, период или момент времени, к которому относится данное значение показателя.

Показателями можно назвать и рейтинги, обобщающие различные свойства каждой единицы совокупности и позволяющие ранжировать их для принятия решений, например, в инвестиционной сфере, в сфере образования и т.д.

В классификации показателей важнейшим является подразделение на абсолютные и относительные, прямые и обратные. Абсолютные показатели служат основой вычисления

разнообразных относительных показателей, получаемых путем соотношения абсолютных величин. Среди абсолютных показателей выделяют число единиц, по которым проводятся расчеты обобщающих показателей, и итоговый подсчет, т.е. суммарное значение какого-либо признака. Значения этих абсолютных показателей определяют степень доверия к относительным и средним показателям.

Относительные показатели подразделяются на характеристики структуры, показатели эффективности и интенсивности производства, сравнительные характеристики (выполнение норм, соответствие нормативу, сравнение с прошлым периодом и т.д., или сравнение разных объектов по одним и тем же показателям за одно и то же время). Особое место в системе статистических показателей занимают средние величины.

Качественный экономический анализ должен быть основан не на отдельных показателях, а на системе показателей, т.е. на группе взаимосвязанных показателей. При этом нужно следовать определенным принципам их построения. Особые сложности возникают, когда показатель должен обобщить разнонаправленные значения (положительные, отрицательные, нулевые).

Основная функция статистических показателей и их систем — познавательно-информационная, однако показатели выполняют и другие функции: прогностическую, оценочную, рекламно-пропагандистскую.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Плошко Б. Г.* Группировка и системы статистических показателей. — М.: Статистика, 1971.
2. *Суслов И. П.* Теория статистических показателей. — М.: Статистика, 1975.
3. *Суслов И. П.* Основы теории достоверности статистических показателей. — Новосибирск: СО «Наука», 1979.

Глава
4

ПРЕДСТАВЛЕНИЕ СТАТИСТИЧЕСКИХ ДАННЫХ: ТАБЛИЦЫ И ГРАФИКИ

4.1. Статистические таблицы

Статистические данные должны быть представлены так, чтобы ими было удобно пользоваться. Существуют по крайней мере три способа представления данных: они могут быть включены в текст, в таблицы или выражены графически.

Если мы включим множество цифр в текст, это затруднит их восприятие. Например, данные об общем числе городов в России и количестве городов с разной численностью населения изложены в следующем тексте.

На 1 января 1998 г. в Российской Федерации было 1090 городов (без Чеченской Республики), из них с численностью населения до 20 тыс. чел. — 379 городов, или 34,8%; городов с численностью населения от 20 до 50 тыс. чел. — 371, или 34,0%; от 50 до 100 тыс. чел. — 176 городов, или 16,2%; от 100 до 500 тыс. чел. — 132 города, или 12,1%; от 500 тыс. чел. до 1 млн чел. — 20 городов, или 1,8%; городов-миллионеров (1 млн жителей и более) — 12, или 1,1%. На 1 января 2002 г. общее число городов составило 1093, из них с численностью населения до 20 тыс. чел. — 402 города, или 36,8%; городов с численностью населения от 29 до 50 тыс. чел. — 357, или 32,7%; от 50 до 100 тыс. чел. — 171 город, или 15,6%; от 100 до 500 тыс. чел. — 132 города, или 12,1%; от 500 тыс. чел. до 1 млн чел. — 21 город, или 1,9%; городов-миллионеров (1 млн жителей и более) — 10, т.е. на 17% меньше, чем в 1998 г.

Даже этот краткий текст из-за перегруженности цифровыми данными плохо воспринимается.

Более эффективно представление статистических данных в форме таблицы.

В отличие от математических таблиц умножения, тригонометрических функций, логарифмов и других, которые по начальным условиям позволяют получить тот или иной результат, статистические таблицы рассказывают языком цифр об изучаемых объектах.

Статистическая таблица — система строк и столбцов, в которых в определенной последовательности и связи излагается статистическая информация о социально-экономических явлениях.

Представим в форме таблицы информацию о городах Российской Федерации (табл. 4.1).

Данные этой таблицы позволяют увидеть незначительный рост общего количества городов при уменьшении числа городов-миллионеров и городов с числом жителей от 20 до 50 тыс. чел., но при увеличении количества городов с численностью

Таблица 4.1

Распределение городов Российской Федерации по численности постоянного населения (по состоянию на 1 января)

Число жителей, тыс. чел.	1998		2002		2002 в % к 1998
	Число городов	%	Число городов	%	
Всего	1090	100	1093	100	100,3
В том числе:					
До 20	379	34,8	402	36,8	106,1
20—49,9	371	34,0	357	32,7	96,2
50—99,9	89	16,2	171	15,6	192,1
100—499,9	132	12,1	132	12,1	100
500—999,9	20	1,8	21	1,9	105
1 млн и более	12	1,1	10	0,9	83,3

Источник. Численность населения Российской Федерации по городам, поселкам городского типа и районам на 1 января 1998 г. — М.: Госкомстат России, 1998. — С. 16—17. Численность населения Российской Федерации по городам, поселкам городского типа и районам на 1 января 2002 г. — М.: Госкомстат России, 2002. — С. 34—35.

населения от 50 до 100 тыс. чел. почти в два раза. В целом изменения свидетельствуют о тенденции выравнивания структуры поселений.

Различают *подлежащее* и *сказуемое* статистической таблицы. В подлежащем указывается характеризуемый объект — либо единицы совокупности, либо группы единиц, либо совокупность в целом. В сказуемом дается характеристика подлежащего, обычно в количественной форме — в виде системы показателей (см. гл. 3). Обязателен *заголовок* таблицы, в котором указывается, к какой категории и какому времени относятся данные таблицы.

По характеру подлежащего статистические таблицы подразделяются на *простые*, *групповые*, *комбинационные*.

В подлежащем простой таблицы объект изучения не подразделяется на группы, а дается либо перечень всех единиц совокупности, либо указывается совокупность в целом. В первом случае таблица называется *простой перечневой*. Единицы упорядочиваются по одному-двум признакам (по возрастанию или убыванию значений). Сказуемое должно содержать данные по каждой единице совокупности. Конечно, построение такой таблицы имеет смысл для принятия каких-то оперативных решений; например, для распределения дополнительных дежурств в больнице нужно знать, сколько дней отработала каждая медсестра за месяц. Такие таблицы хороши при небольшом числе единиц (20 и менее). Скажем, подобную таблицу можно построить для характеристики работы метрополитена в городах России, так как метро имеется лишь в пяти городах.

При большом (несколько десятков и более) числе единиц простые перечневые таблицы составляются только как вспомогательные, например как основа последующей группировки.

Простые таблицы, содержащие данные о совокупности в целом, можно встретить очень часто в газетах, статистических сборниках. Как правило, они представляют данные в динамике. Примером такой таблицы является табл. 4.2, в которой приведена структура макроэкономического показателя — использованного валового внутреннего продукта России.

В подлежащем *групповой таблицы* объект изучения подразделяется на группы по *одному признаку*. В сказуемом указываются число единиц в группах (абсолютное и в процентах

Таблица 4.2

Использование валового внутреннего продукта Российской Федерации (в процентах к итогу в фактически действующих ценах)

Год	ВВП использованный	В том числе		
		расходы на конечное потребление	валовое накопление	чистый экспорт товаров и услуг
1993	100	64,2	27,8	8,0
2000	100	62,5	17,1	20,4

Источник. Национальные счета России в 1993—2000 годах. Статистический сборник. — М.: Госкомстат России, 2001. — С. 57.

к итогу) и сводные показатели по группам. Примером такой таблицы является табл. 4.3, в которой изучаемая совокупность — занятое население России — распределяется по формам собственности. Табл. 4.1 также является групповой.

В подлежащем *комбинационной таблицы* совокупность подразделяется на группы не по одному, а по *нескольким признакам*. Например, в табл. 4.4 изучаемая совокупность — занятое население России — подразделяется на группы по двум признакам: возрасту и полу.

Кроме перечисленных может использоваться типовая таблица, в подлежащем которой дается словесная характеристика

Таблица 4.3

Среднегодовая численность занятых в экономике по формам собственности, млн чел.

	1992	2001
Всего	72,1	65,0
В том числе по формам собственности:		
государственная, муниципальная	49,7	24,3
частная	14,0	31,1
собственность общественных и религиозных организаций (объединений)	0,6	0,5
смешанная российская	7,6	7,1
иностранная, совместная российская и иностранная	0,2	2,0

Источник. Россия в цифрах. 2002. Краткий статистический сборник. — М.: Госкомстат России, 2002. — С. 79.

Таблица 4.4

Распределение численности занятых в экономике в 2001 г. по возрастным группам (на конец ноября, в процентах к итогу)

Возрастная группа, лет	Всего	Мужчины	Женщины
Всего	100	100	100
В том числе в возрасте:			
до 20	2,1	2,4	1,7
20—24	9,6	10,2	8,9
25—29	12,5	13,1	11,8
30—34	11,7	12,2	11,1
35—39	14,3	14,2	14,5
40—44	16,1	15,2	17,2
45—49	14,6	13,6	15,7
50—54	11,1	10,4	11,9
55—59	3,8	4,3	3,3
60—72	4,2	4,4	3,9

Источник. Россия в цифрах. 2002. Краткий статистический сборник. — М.: Госкомстат России, 2002. — С. 84.

выделенных типов, но как они получены, с помощью каких группировочных признаков, это в таблице не указывается. Можно сказать, что в типовой таблице обобщаются ранее принятые решения о группировке изучаемой совокупности. Например, рабочие могут подразделяться на низкоквалифицированных и высококвалифицированных, экономически активное население — на занятых и безработных, города — на малые, средние и крупные. Но как определялся уровень квалификации рабочего или кого отнесли к безработным, или какие города отнесены к малым, средним, крупным, это из таблицы не следует. В ней представлен результат решений, принятых ранее.

При построении таблиц необходимо руководствоваться следующими общими правилами.

Подлежащее таблицы располагается в левой части, сказуемое — в правой, но могут быть исключения. В простой таблице (см. табл. 4.2) подлежащее, т.е. объект изучения, указано в заголовке таблицы; в комбинационной таблице подлежащее

может располагаться в левой и верхней частях таблицы (см. табл. 4.4).

В таблице не должно быть ни одной лишней линии, только необходимые: линия, отделяющая заголовок таблицы от заголовков ее граф, заголовки граф от цифровых данных. Иногда используется линия, отделяющая итоговую строку. Вертикальная разграфка может быть, а может и отсутствовать.

Заголовки граф содержат названия показателей (без сокращения слов), их единицы измерения. Последние могут указываться как в заголовке соответствующей графы, так и в заголовке таблицы или над таблицей (см., например, табл. 4.4), если все показатели таблицы выражены в одних и тех же единицах измерения и счета.

Итоговая строка завершает таблицу и располагается в конце таблицы, но иногда бывает первой: в этом случае во второй строке дается запись «В том числе», и последующие строки содержат составляющие итоговой строки, иногда не все, а основные.

Цифровые данные записываются с одной и той же степенью точности в пределах каждой графы: при этом обязательно разряды чисел располагаются под разрядами, целая часть числа отделяется от дробной запятой, например 4,5, а не 4.5. Заметим, что в международных статистических публикациях вместо запятой используется точка; цифры целой части числа в два раза больше дробной 4.5. В таблице не должно быть ни одной пустой клетки: если данные равны нулю, ставится знак «—» (прочерк); если данные не известны, делается запись «сведений нет» или ставится знак «...» (троеточие). Если значение показателя не равно нулю, но первая значащая цифра появляется после принятой степени точности, то делается запись 0,0 (если, скажем, была принята степень точности 0,1).

Если таблица имеет много граф, то графы подлежащего обозначаются заглавными буквами (А, Б), а графы сказуемого — цифрами (1, 2 и т.д.). Это бывает удобно; если таблица имеет много строк и печатается на нескольких страницах, то заголовки граф не повторяются, а указываются только их обозначения.

Если таблица основана на заимствованных данных, то под ней указывается источник данных (см., например, табл. 4.2).

Если хотите, чтобы построенная вами таблица была понятна и удобна для пользования, не пренебрегайте ни одним из указанных правил.

Теперь вам помогает строить таблицы Excel, и программное обеспечение избавляет вас от многих забот. Тем не менее ответственность за подготовку статистических таблиц возлагается на статистика, а не на компьютер.

4.2. Основные виды графиков

Статистические таблицы дополняются графиками в том случае, когда ставится цель подчеркнуть какую-то особенность данных, провести их сравнение. Графики являются самой эффективной формой представления данных с точки зрения восприятия. Часто графики используются и вне связи с таблицей. С помощью графиков достигается наглядность характеристики структуры, динамики, взаимосвязи явлений, их сравнения.

Статистические графики представляют собой условные изображения числовых величин и их соотношений посредством линий, геометрических фигур, рисунков или географических карт-схем.

Графический способ облегчает рассмотрение статистических данных. На графике сразу видны пределы изменения показателя, сравнительная скорость изменения разных показателей, их колеблемость. Вместе с тем график имеет определенные ограничения: прежде всего не может включить столько данных, сколько может войти в таблицу; кроме того, на нем показываются всегда округленные данные — не точные, а приблизительные. Таким образом, график используется только для изображения общей ситуации, а не деталей. Последний минус — трудоемкость построения. Но этот недостаток может быть преодолен применением пакетов прикладных программ (ППП) для компьютерной графики, например ППП «Harvard graphics».

По способу построения графики делятся на *диаграммы*, *картограммы* и *картодиаграммы*.

Наиболее распространенными являются диаграммы. Они бывают разных видов: линейные, радиальные, точечные, плоскостные, объемные, фигурные. Вид диаграммы зависит от вида представляемых данных (одна переменная или один показатель, несколько переменных или показателей, количественные или неколичественные) и задачи построения графика.



Рис. 4.1. Динамика выбросов вредных веществ в атмосферу и индекса физического объема промышленного производства в Санкт-Петербурге

В любом случае график обязательно сопровождается заголовком — над или под полем графика. В заголовке указывается, какой показатель изображен, в каких единицах измерения, по какой территории и за какое время он определен.

Линейные графики используются для представления количественных переменных: характеристики вариации их значений, динамики, взаимосвязи между переменными.

Вариация данных анализируется с помощью полигона распределения, *кумуляты* (кривой «не меньше, чем») и *огивы* (кривой «больше, чем»). Линейные графики используются в решении задач классификации данных. Линейные графики применяются в анализе динамики связей. В анализе используются точечные диаграммы (так называемое поле корреляции).

Линейные графики целесообразно разделять на используемые для представления данных *по одной переменной* — одномерные или *по двум переменным* — двумерные. Примером первого является полигон распределения, второго — линия регрессии. Возможен такой случай, когда на графике пред-

ставлено несколько переменных (показателей), а он все-таки не является многомерным (рис. 4.1).

Для того чтобы динамика двух и более показателей была сопоставимой, следует обеспечить их «единый старт», как на рис. 4.1, где показатели 1990 г. приняты за 100%.

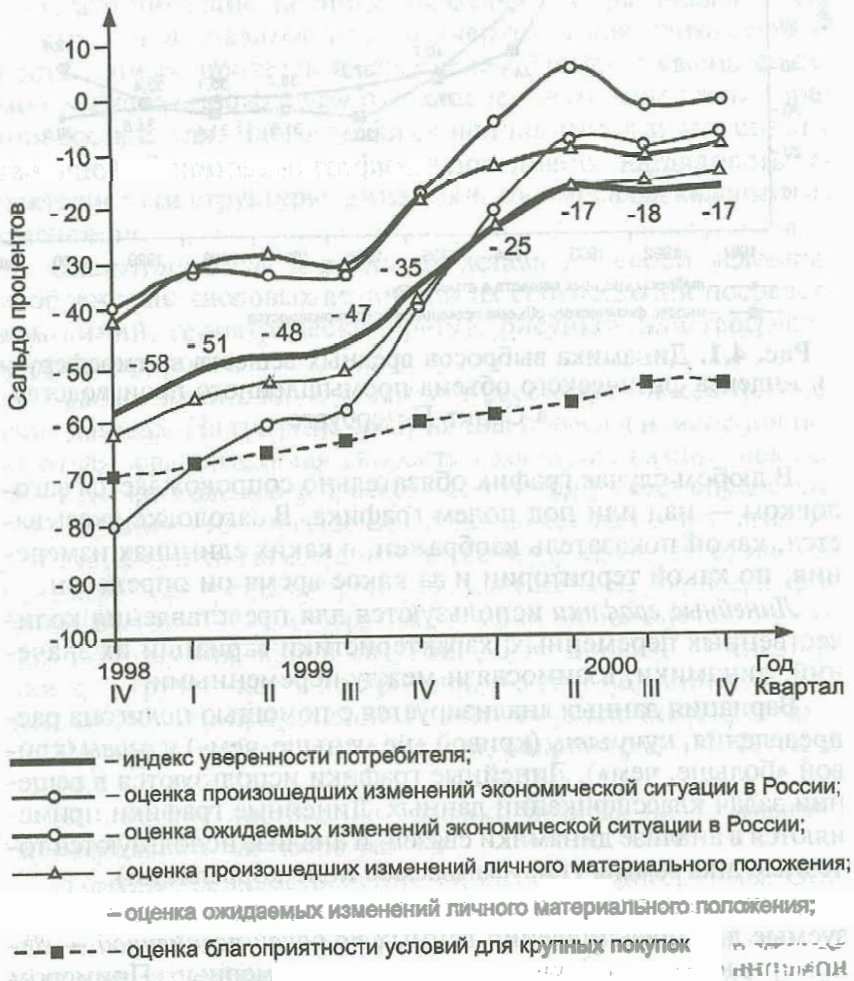


Рис. 4.2. Индекс уверенности потребителя (I кв. — февраль, II кв. — май, III кв. — август, IV кв. — ноябрь)

Динамика двух показателей на одном и том же графике может быть представлена и без приведения их к 100%, если эти показатели связаны каким-либо функциональным соотношением (например, представлена динамика общего показателя и показателя, который является одним из его составляющих). Примером такого графика является рис. 4.2.

При графическом изображении динамики по оси абсцисс показывается время (годы, кварталы, месяцы); по оси ординат — значения показателей или показателя (рис. 4.3, *а*). При этом ось ординат должна иметь начало в точке «0». Иногда вместо нулевой точки в качестве начального уровня на оси ординат показывается уровень какого-либо года. Это делается в том случае, если изменения изображаемого показателя значительны — в 8—10 раз и более в течение рассматриваемого отрезка времени. Однако такой прием не рекомендуется. Правильнее указать нулевую точку, а затем (если нужно) «разорвать» ось ординат так, как это показано на рис. 4.3, *б*.

Иногда при больших изменениях показателя прибегают к логарифмической шкале. Предположим, значения показателя изменяются от 1 до 100 (в 100 раз); это может вызвать затруднения при построении графика. Если перейти к логарифмам, то их значения для минимальных (максимальных) значений показателя будут различаться не так сильно: $\log 1 = 0$, $\log 100 = 2$.

Среди *плоскостных диаграмм* по частоте использования выделяются *столбиковые диаграммы*, на которых показатель представляется в виде столбика, высота которого соответствует значению показателя. Пример столбиковой диаграммы представлен на рис. 4.4. Часто на столбиковой диаграмме показываются относительные величины: при сравнении показателей по группам, по разным совокупностям, одна из которых может быть принята за 100%.

Пропорциональность площади той или иной геометрической фигуры величине показателя лежит в основе других видов плоскостных диаграмм: *треугольных*, *квадратных*, *прямоугольных*. В треугольной диаграмме нужно так выбрать стороны и высоту треугольника, чтобы его площадь отвечала величине показателя. Для построения квадратной диаграммы нужно задать размер одной стороны, прямоугольной — двух

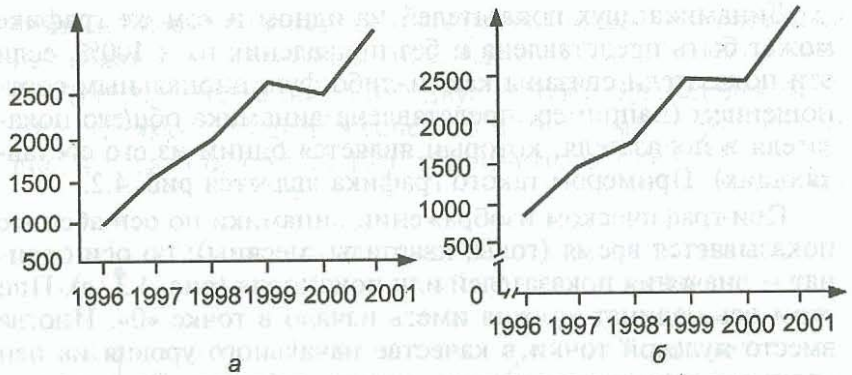


Рис. 4.3. Включение нулевой точки при изображении динамики: *а* — неверно; *б* — верно

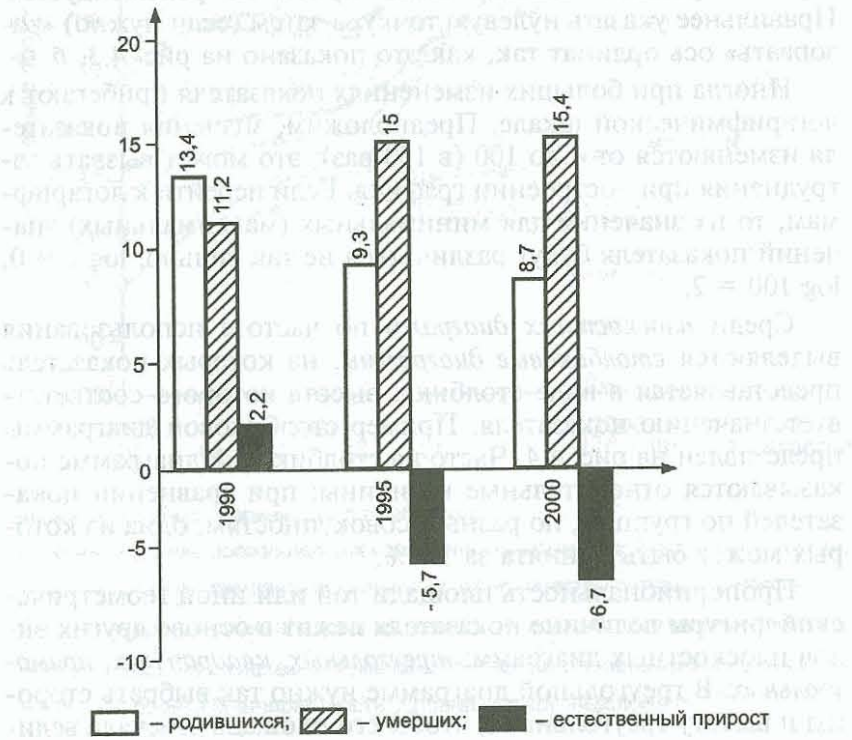


Рис. 4.4. Общие показатели рождаемости, смертности и естественного прироста населения России, в ‰

сторон. Можно использовать и сравнение площадей круга; в этом случае задается радиус окружности.

Ленточная диаграмма представляет показатели в виде горизонтально вытянутых прямоугольников. Как столбиковые, так и ленточные диаграммы можно применять не только для сравнения самих величин, но и для сравнения их частей (рис. 4.5 и 4.6).

Особый тип ленточных диаграмм применяется для представления данных с разным характером изменений: положительным и отрицательным (рис. 4.7).

Диаграмма, изображенная на рис. 4.7, может использоваться, например, для представления регионов с разной величиной и характером миграционного сальдо (положительным и отрицательным) предприятий, на которых повысилась и понизилась оплата труда и т.д.

Из плоскостных диаграмм часто используется *секторная диаграмма*. Она применяется для иллюстрации структуры изучаемой совокупности. Вся совокупность принимается за

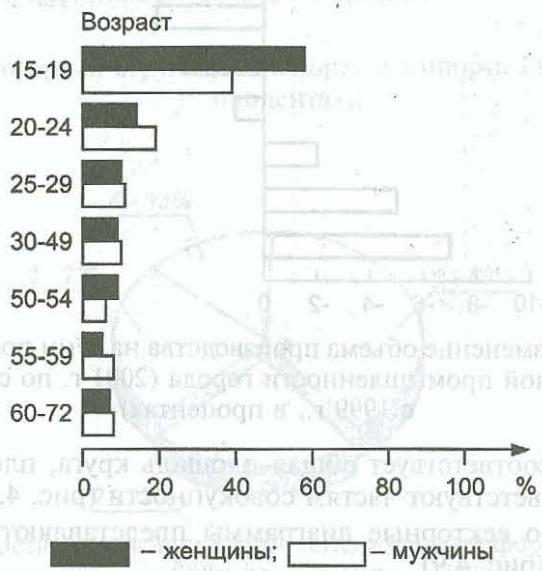


Рис. 4.5. Доля безработных в экономически активном населении крупного города

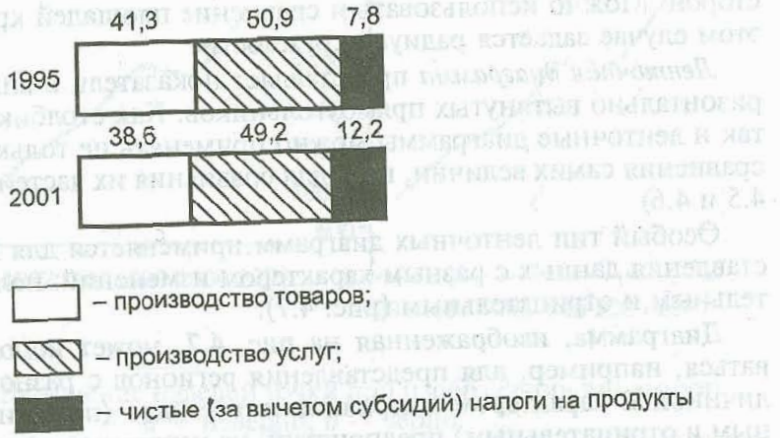


Рис. 4.6. Структура производства валового внутреннего продукта РФ (в текущих ценах; в процентах к итогу)

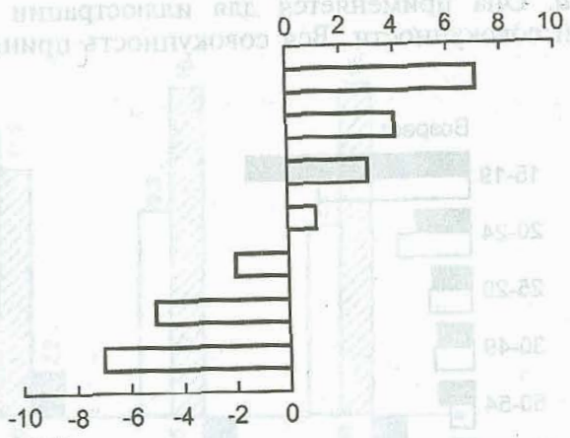
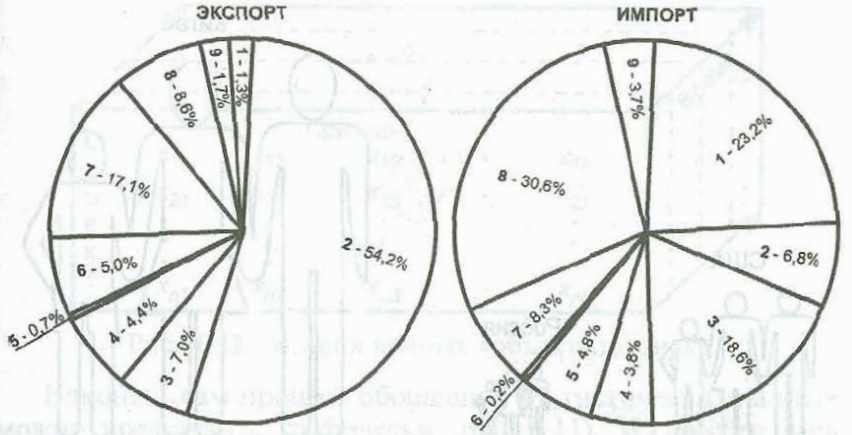


Рис. 4.7. Изменение объема производства на семи предприятиях текстильной промышленности города (2001 г. по сравнению с 1999 г., в процентах)

100%, ей соответствует общая площадь круга, площади секторов соответствуют частям совокупности (рис. 4.8).

Нередко секторные диаграммы представляют в следующем виде (рис. 4.9).

Фигурные, или картинные, *диаграммы* усиливают наглядность изображения, так как включают рисунок изображаемо-



1 – продовольственные товары и сельскохозяйственное сырье;
2 – минеральные продукты;
3 – продукция химической промышленности, каучук;
4 – древесина и целлюлозно-бумажные изделия;
5 – текстиль, текстильные изделия и обувь;
6 – драгоценные камни, драгоценные металлы и изделия из них;
7 – металлы и изделия из них;
8 – машины, оборудование и транспортные средства;
9 – прочие

Рис. 4.8. Товарная структура экспорта и импорта РФ в 2000 г. (в процентах):

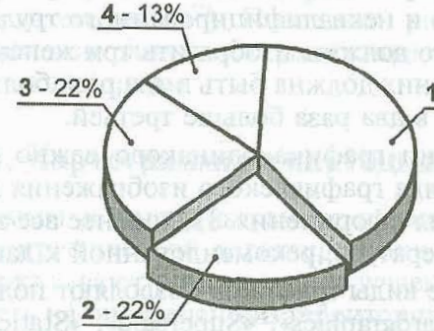


Рис. 4.9. Состав российских топ-менеджеров по форме участия в бизнесе в 2001 г.:
1 — наемные менеджеры; 2 — собственники;
3 — владельцы менеджерского пакета; 4 — другие

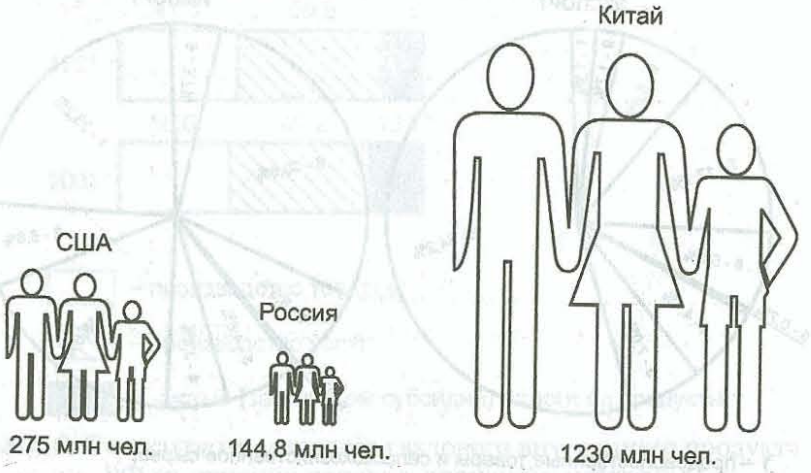


Рис. 4.10. Численность постоянного населения на конец 2000 г.

го показателя. Площадь фигуры соответствует величине показателя (рис. 4.10).

Если, например, вы решите использовать фигурную диаграмму для изображения структуры безработных женщин, среди которых 47% — молодые женщины (20—24 года) и девушки 16—19 лет, не имеющие стажа работы; 28% — инженерно-технические работники и служащие со специальным образованием в возрасте 25—49 лет и 15% — работницы квалифицированного и неквалифицированного труда в возрасте 50 лет и старше, то должны изобразить три женские фигуры, причем первая из них должна быть в два раза больше второй, а вторая — почти в два раза больше третьей.

При построении графика одинаково важно все — правильный выбор вида графического изображения пропорций, соблюдение правил оформления. Подробнее все эти вопросы освещаются в литературе, рекомендованной к данной главе.

Разнообразные виды графиков позволяют получить ППП для ПЭВМ «Harvardgraphics», «Supercalc», «Statictica», «Statgraphics» и др. На графическом представлении основаны некоторые процедуры классификации (группировки) данных, анализа динамики: выявление тенденции, сравнение динамики разных показателей и т.д.



Рис. 4.11. Массив данных «объект-признак»

Наконец, сам процесс обобщения статистических данных можно представить графически (рис. 4.11). Изображен весь массив собранных данных, т.е. таблица «объект-признак», полученная за ряд периодов. Например, собраны данные по промышленным предприятиям на данной территории по многим характеристикам за каждый месяц. Это можно представить в виде параллелепипеда, что и сделано на рис. 4.11.

Третье измерение может быть не временем, а определенной территорией, т.е. каждая таблица «объект-признак» относится к определенной территории (району, области и т.д.). На последующих рисунках показано, что каждый из подмассивов, взятых из рис. 4.12, *а*, может выделяться и разрабатываться самостоятельно (*б*); на рис. 4.12, *в* и *г* показано, что данные могут подразделяться по регионам, по кварталам и, наконец, по категориям (*д*). Последний рис. 4.12, *е* изображает подразделение данных по трем основаниям: по времени, территории и категориям.

4.3. Картограммы и картодиаграммы

Картограммы и картодиаграммы применяются для изображения географической характеристики изучаемых явлений. Они показывают размещение изучаемого явления, его интенсивность на определенной территории — в республике, области, экономическом или административном районе и т.д.

На *картограмме* распределение изучаемого признака по территории изображается условными знаками (точками, штриховкой, цветом и т.д.), соответствующими определен-

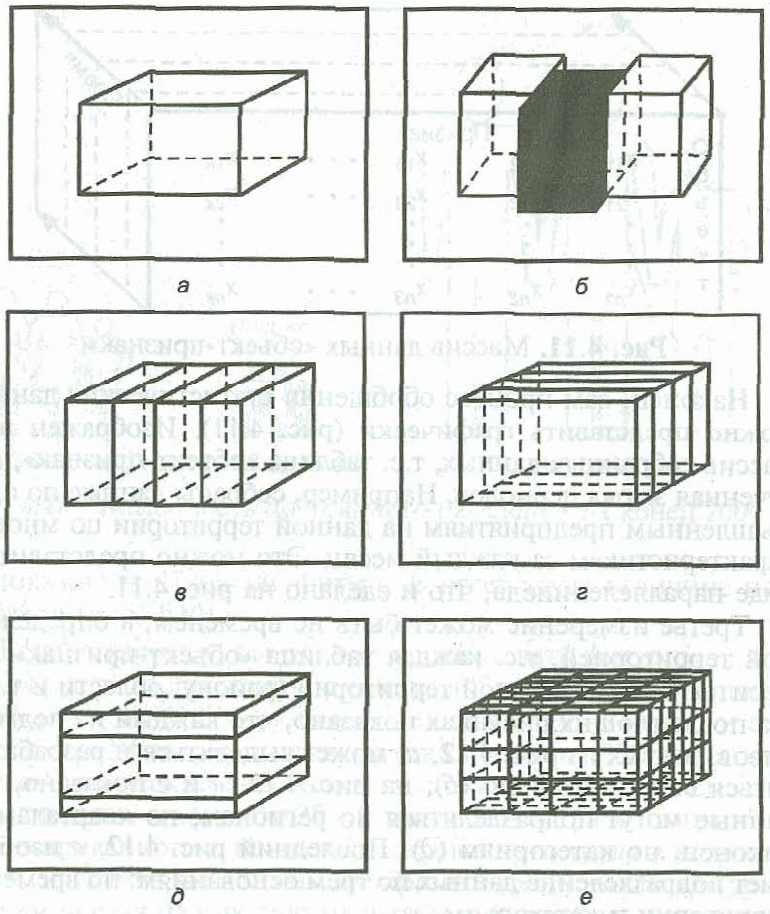


Рис. 4.12. Обработка статистических данных:
а — исходные данные; *б* — выделение подмассива; *в* — выделение по кварталам; *г* — подразделение по регионам; *д* — подразделение по категориям; *е* — подразделение по трем признакам

ным интервалам значений величины этого признака. Эти знаки покрывают контур каждого района. Картограмма применяется в тех случаях, когда возникает необходимость показать территориальное распределение какого-нибудь одного статистического признака между отдельными районами для выявления закономерностей этого распределения.

Картограммы бывают *фоновые* и *точечные*. На фоновых картограммах распределение изучаемого явления на территории изображается различной раскраской территориальных единиц с разной густотой цвета. Часто вместо раскраски применяется штриховка различной интенсивности. Такие картограммы обычно используются для изображения уровня относительных и средних величин по территориям. Например, имеются данные об урожайности зерновых по 10 районам области: урожайность до 20 ц/га имеют три смежных района, 20—30 ц/га — четыре смежных района, свыше 30 ц/га — три смежных района. Соответствующая фоновая картограмма представлена на рис. 4.13. Чем интенсивнее явление, тем гуще штриховка (точки), или темнее окраска. Такая картограмма наглядно показывает географию урожайности зерновых культур по районам. Чем больше групп, тем точнее изображение, но большое число групп создает пестроту и снижает наглядность. Поэтому лучше всего применять не более четырех-пяти тонов, или градаций плотности штриховки.

На точечной картограмме символами графического изображения статистических данных являются точки, размещенные в пределах определенных территориальных границ. Точечная картограмма применяется для изображения абсолютных величин. Каждой точке, нанесенной на картограмму, придается числовое значение, что позволяет использовать ее для прямого

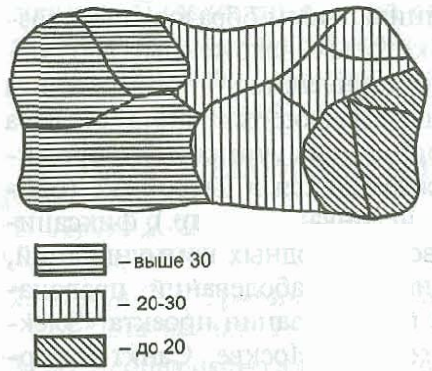


Рис. 4.13. Картограмма распределения районов по урожайности зерновых, ц/га

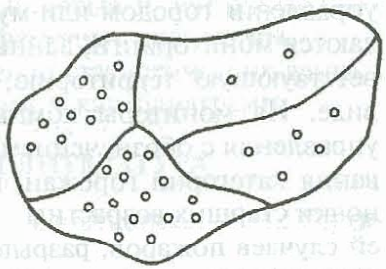


Рис. 4.14. Картограмма добычи угля по районам

счета. Например, имеются четыре района с добычей угля в 200, 500, 1000 и 1400 тыс. т в год. Для составления картограммы примем точку за 100 тыс. т и нанесем на контур каждого района соответствующее количество точек (рис. 4.14).

Картодиаграмма — это сочетание диаграммы с географической картой. В качестве изобразительных знаков в картодиаграммах используются те или иные фигуры, которые размещаются на контуре географической карты. Картодиаграммы дают возможность графически отразить более сложные статистико-географические соотношения, чем картограммы. Так, с помощью картодиаграммы можно выразить пространственную специфику в структурах изучаемых статистических совокупностей, особенности каждого района как единого целого и т.д. Например, структурная, или секторная, картодиаграмма, характеризующая порайонные различия в структуре посевных площадей. В качестве диаграммных знаков в картодиаграмме часто используют различные геометрические фигуры, особенно круги, которые наиболее просты и удобны для выражения сравниваемых количественных показателей.

Кроме рассмотренных видов диаграмм, картограмм и картодиаграмм на практике встречаются и другие, более сложные графические изображения статистических данных.

В настоящее время пространственное представление статистических данных используется все шире. Например, для управления городом или муниципальным образованием создаются мониторинги, данные которых, наложенные на соответствующую территорию, представляются в электронном виде. На мониторы компьютера выводится карта объекта управления с обозначением социальных учреждений, проживания категорий горожан, нуждающихся в патронаже (одиночки старших возрастных групп, инвалиды и др.), фиксацией случаев пожаров, разрыва водопроводных коммуникаций, канализации, случаев инфекционных заболеваний, правонарушений и пр. Ведется работа по реализации проекта «Электронная Россия». В крупных городах Москве, Санкт-Петербурге — на базе данных Всероссийской переписи населения 2002 г. реализуются проекты «Электронная Москва», «Электронный Санкт-Петербург».

РЕЗЮМЕ

Наиболее удобная и рациональная форма представления количественных данных — таблица. Статистическая таблица должна быть построена по определенным правилам. Она состоит из подлежащего (объект изучения) и сказуемого (цифровая характеристика объекта).

Вид таблицы определяется по подлежащему — по тому, как представлен объект изучения:

- простая таблица — объект изучения не разделен на группы, т.е. показываются либо единицы совокупности, либо совокупность в целом;
- групповая таблица — объект изучения разделен на группы по одному признаку;
- комбинационная таблица — объект изучения разделен на группы по двум и более признакам;
- типовая таблица — объект изучения разделен на типы, и в подлежащем дана словесная характеристика типов.

Сказуемое также должно оформляться по правилам. Использование программного обеспечения Excel позволяет обеспечить качество построения статистических таблиц.

Таблица должна иметь заголовок; должен быть указан источник данных.

Графики обеспечивают наглядность представления данных; они подразделяются на линейные, плоскостные и секторные. Плоскостные — на столбиковые и ленточные диаграммы. Широко используются фигурные диаграммы.

Пространственное представление статистических данных достигается с помощью картограмм и картодиаграмм.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Герчук Я. П.* Графические методы в статистике. — М.: Статистика, 1968.
2. *Герчук Я. П.* Графики в математико-статистическом анализе. — М.: Статистика, 1972.
3. *Курс* социально-экономической статистики / Под ред. М. Г. Назарова. — М.: Финстатинформ, 2002.
4. *Теория* статистики / Под ред. Р. А. Шмойловой. — 4-е изд., доп. и перераб. — М.: Финансы и статистика, 2003.

Глава
5

СРЕДНИЕ ВЕЛИЧИНЫ И ИЗУЧЕНИЕ ВАРИАЦИИ

5.1. Однородность и вариация массовых явлений

Как уже было сказано, статистика изучает массовые явления и процессы. Каждое из таких явлений обладает как общими для всей совокупности, так и особенными, индивидуальными свойствами. Различие между индивидуальными явлениями называют *вариацией* (подразд. 5.5). Здесь же рассмотрим другое свойство массовых явлений — присущую им близость характеристик отдельных явлений. Если в сосуд с горячей водой добавить холодную, то температура воды во всем сосуде станет одинаковой (осреднится). Поведение детей, поступивших в одну группу детского садика или в один класс школы, тоже приобретает до какой-то степени общие, усредненные черты. Массовое промышленное производство невозможно без стандартизации, т.е. усреднения размеров деталей собираемых механизмов, узлов, агрегатов. Введение севооборота, т.е. ротация разных культур по нескольким участкам пашни, приведет к выравниванию плодородия и механических свойств почвы на этих севооборотных полях. Итак, взаимодействие элементов совокупности приводит к ограничению вариации хотя бы части их свойств. Эта тенденция существует объективно. Именно в ее объективности заключена причина широчайшего применения средних величин в теории и на практике.

Каждому рабочему известно, что оплата за простой не по вине рабочего проводится по средним расценкам или по среднечасовому заработку. Каждому студенту известно, что такое средний балл на экзаменах. О средних величинах и серьезно,

и с насмешкой говорят и пишут философы и журналисты. С помощью метода средних величин статистика решает много задач.

Главное значение средних величин состоит в их *обобщающей функции*, т.е. замене множества различных индивидуальных значений признака средней величиной, характеризующей всю совокупность явлений. Всем известны особенности развития современных людей, проявляющиеся в том числе и в более высоком росте сыновей по сравнению с отцами, дочерей в сравнении с матерями в том же возрасте. Но как измерить это явление? В разных семьях наблюдаются самые различные соотношения роста старшего и младшего поколения. Далеко не всякий сын выше отца и не каждая дочь выше матери. Но если измерить средний рост многих тысяч лиц, то по среднему росту сыновей и отцов, дочерей и матерей можно точно установить и сам факт акселерации, и типичную среднюю величину увеличения роста за одно поколение.

На производство одного и того же количества товара определенного вида и качества разные производители (заводы, фирмы) затрачивают неодинаковое количество труда и материальных ресурсов. Но рынок осредняет эти затраты, и стоимость товара определяется средним расходом ресурсов на производство.

Погода в определенном пункте земного шара в один и тот же день в разные годы может быть очень различной. Например, в Санкт-Петербурге 31 марта температура воздуха за сто с лишним лет наблюдений колебалась от −20,1° в 1883 г. до +12,24° в 1920 г. Примерно такие же колебания наблюдаются и в другие дни года. По таким индивидуальным данным о погоде в какой-то произвольно взятый год нельзя составить представление о климате Санкт-Петербурга. Характеристики климата — это средние за длительный период характеристики погоды — температура воздуха, его влажность, скорость ветра, сумма осадков, число часов солнечного сияния за неделю, месяц и весь год и т.д. Приведем еще один пример осреднения, его роли в управлении важнейшими и опасными процессами, от которых зависит жизнь людей. Физика установила, что невозможно предсказать, когда произойдет распад ядра радиоактивного атома, например изотопа уран-235. Атом может распасться через секунду или через тысячу лет. Но в

массе атомов (например, находящихся в стержнях реактора АЭС) точно можно измерить среднюю скорость распада (обычно используют показатель «время полураспада» — время, за которое распадается половина атомов). Вводя вещества-замедлители образующихся при распаде атомов урана частиц или убирая их, можно управлять скоростью цепной реакции в урановых стержнях, регулировать мощность реактора, вводить ее в безопасные и экономически выгодные режимы.

Если средняя величина обобщает качественно однородные значения признака, то она является типической характеристикой признака в данной совокупности. Так, можно говорить об измерении типичного роста русских девушек рождения 1983 г. по достижении ими 20-летнего возраста. Типичной характеристикой будет средняя величина надоя молока от коров черно-пестрой породы на первом году лактации при норме кормления 12,5 кормовой единицы в сутки. Для лиц с достаточно однородным уровнем дохода, например рабочих машиностроительной отрасли, пенсионеров по старости (исключая имеющих льготы), можно определить типичные доли расходов на покупку предметов питания в их бюджете.

Однако неправильно сводить роль средних величин только к характеристике типичных значений признаков в однородных по данному признаку совокупностях. На практике современная статистика значительно чаще использует средние величины, обобщающие явно неоднородные явления, как, например, урожайность всех зерновых культур по территории всей России, включая кукурузу, дающую по 50—60 ц/га и более, и гречиху, дающую 6—10 ц/га, и плодородные черноземы Кубани, и скудные почвы Архангельской области. Или рассмотрим такую среднюю, как среднее потребление мяса на душу населения: ведь среди этого населения и дети до одного года, вовсе не потребляющие мяса, и вегетарианцы, и северяне, и южане, шахтеры, спортсмены и пенсионеры. Еще более ясна нетипичность такого среднего показателя, как произведенный национальный доход в среднем на душу населения.

Средняя величина национального дохода на душу, средняя урожайность зерновых по всей стране, среднее потребление разных продуктов питания — это характеристики государства как единой народнохозяйственной системы, так называемые *системные средние.*

Системные средние могут характеризовать как пространственные, или объектные, системы, существующие одномоментно (государство, отрасль, регион, планета Земля и т.п.), так и динамические системы, протяженные во времени (год, десятилетие, сезон и т.п.). Примером системной средней, характеризующей период времени, может служить средняя температура воздуха в Санкт-Петербурге за 1996 г., равная +5,19 °С. Эта средняя величина обобщает и летние высокие температуры +20, +25°, и зимние морозы, осень и весну, дни и ночи.

С другой стороны, средняя температура воздуха за отдельный год не является типической характеристикой климата Санкт-Петербурга, потому что в разные годы средняя температура года значительно колеблется, например за последние 30 лет от +2,90° в 1976 г. до +7,44° в 1989 г. Типической характеристикой климата будет многолетняя средняя годовая температура за десятки лет, например, за 1967—1996 гг. она составила +5,05°.

Итак, типическая средняя может обобщать системные средние для однородной совокупности, или системная средняя может обобщать типические средние для единой, хотя и неоднородной, системы. При этом даже типическая средняя не является раз и навсегда данной, неизменной характеристикой.

Так, многолетняя средняя температура в Санкт-Петербурге в первые десятилетия и столетие существования города была значительно ниже; она возрастает медленно, но с ускорением за последнее столетие вследствие как роста самого города и энергопотребления в нем, что повышает температуру воздуха, так и начавшегося, и ускоряющегося общего потепления на земле. Поэтому «типичность» любой средней величины — понятие относительное, ограниченное как в пространстве, так и во времени.

5.2. Средняя арифметическая величина

Понятие средней арифметической

Виды средних величин различаются прежде всего тем, какое свойство, какой параметр исходной варьирующей массы индивидуальных значений признака должен быть сохранен неизменным.

Средней арифметической величиной называется такое значение признака в расчете на единицу совокупности, при вычислении которого общий объем признака в совокупности сохраняется неизменным.

Иными словами, средняя арифметическая величина — среднее слагаемое. При ее вычислении общий объем признака мысленно распределяется поровну между всеми единицами совокупности. Например, средняя заработная плата, или средний доход, работников предприятия — это такая сумма денег, которая приходилась бы на каждого работника, если бы весь фонд оплаты труда (или все доходы, направленные на личное потребление) был распределен между работниками поровну.

Исходя из определения формула средней арифметической величины имеет вид:

$$\bar{x} = (x_1 + x_2 + \ldots + x_n) : n = \frac{\sum_{i=1}^{n} x_i}{n}, \quad (5.1)$$

$\bar{x}$ — средняя величина;

n — численность совокупности.

По формуле (5.1) вычисляются средние величины *первичных (объемных) признаков*, если известны индивидуальные значения признака. Если изучаемая совокупность велика, исходная информация чаще представляет собой ряд распределения, или группировку, как, например, табл. 5.1.

Среднее число мячей, забитых за одну игру, должно представлять собой результат равномерного распределения общего числа забитых мячей по всем 240 матчам розыгрыша первенства. Общее число забитых мячей согласно табл. 5.1 можно получить как сумму произведений значений признака в

Таблица 5.1

Распределение футбольных матчей высшей лиги России по числу забитых за матч мячей обеими командами в 1999 г.

Число забитых мячей, x_i	0	1	2	3	4	5	6	7	8	9	10	Итого
Число матчей, f_i	21	46	53	51	34	16	14	4	—	—	1	240

каждой группе x_i на число игр с таким количеством забитых мячей f_i (частоты). Получим формулу

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}, \quad (5.2)$$

где n — число групп.

Такую форму средней арифметической величины называют *взвешенной арифметической средней* в отличие от простой средней, рассчитанной по формуле (5.1). В качестве весов здесь выступают числа единиц совокупности в разных группах. Название «вес» отражает тот факт, что разные значения признака имеют неодинаковую «важность» при расчете средней величины. «Важнее», весомее число забитых мячей, которое встречалось чаще: 1, 2, 3 мяча, а такие значения, как 7 или 10 забитых мячей, как бы ни радовались таким результативным матчам болельщики, при расчете средней не играют большой роли: их «вес» мал.

Имеем: $\bar{x} = 2{,}68$ мяча за игру.

Как видим, средняя арифметическая величина может быть дробным числом, если даже индивидуальные значения признака принимают только целые значения (дискретный признак). Ничего «предосудительного» для метода средних в этом не заключено; из сущности средней не вытекает, что она обязана быть реальным значением признака, которое могло бы встретиться у какой-либо единицы совокупности.

Виды средней арифметической

Если при группировке значения осредняемого признака заданы интервалами, то при расчете средней арифметической величины в качестве значения признака в группах принимают середины этих интервалов, т.е. исходят из гипотезы о равномерном распределении единиц совокупности по интервалу значений признака. Для открытых интервалов в первой и последней группе, если таковые есть, значения признака надо определить экспертным путем исходя из сущности, свойств признака и совокупности. Например, по табл. 5.2 можно ми-

Таблица 5.2

Распределение рабочих предприятия по возрасту

Группы рабочих по возрасту, лет	Число рабочих, f_j	Середина интервала, x'	$x_j f_j$
А	1	2	3
До 20	48	18,5	888
20—30	120	25	3000
30—40	75	35	2625
40—50	62	45	2790
Старше 50	54	57,5	3105
Итого	359	34,56	12408

нимальный возраст рабочих считать 17 лет. В таком случае первый интервал будет от 17 до 20 лет, а максимальный возраст — 65 лет, тогда последний интервал — 50—65 лет.

Средний возраст рабочих, рассчитанный по формуле (5.2) с заменой точных значений признака в группах серединами интервалов, составил:

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j' f_j}{\sum_{j=1}^{k} f_j} = \frac{12\,408}{359} = 34{,}56 \text{ года},$$

что и записано в итоговую строку по графе 3 табл. 5.2. Напомним, что итог объемного показателя — это сумма, итог по графе относительных показателей или средних групповых величин — относительная или средняя средняя. Числитель дроби — это общая сумма человеко-лет, прожитых рабочими предприятия; разделив ее на число работников, получаем возраст в годах, так что логика показателя средней величины соблюдена.

Перейдем к рассмотрению средних *вторичных* (относительных) *признаков*. Сумма таких показателей сама по себе реальной величиной какого-либо признака в совокупности не является. Однако общее определение арифметической средней сохраняет силу и в этом случае. При вычислении таких *средних величин необходимо, чтобы сохранялась сумма величины объемного признака, который является числителем при построе-*

нии осредняемого относительного показателя. Например, при вычислении средней величины урожайности какой-либо сельскохозяйственной культуры (по формуле (5.2)) необходимо, чтобы общий объем валового сбора этой культуры остался неизменным при замене индивидуальных величин урожайности средней величиной. Нельзя менять реальную величину объемного признака — она является базой расчета средней. Чтобы выполнить указанное условие, в качестве весов при расчете средней величины относительного показателя необходимо принять значения того признака, который является знаменателем при определении относительного показателя. Так, при вычислении средней урожайности по совокупности хозяйств весами должны служить размеры площади данной культуры.

Пример. Рассчитаем среднюю долю товаров народного потребления в общем выпуске промышленной продукции по совокупности предприятий (табл. 5.3). В этом случае весом должен являться общий объем всей продукции предприятия.

Тогда средняя доля предметов народного потребления в продукции четырех предприятий равна: $\bar{x} = (615{,}5 : 2047) \cdot 100\% = 30{,}07\%$. Средняя доля ближе к значениям долей тех предприятий, которые имеют большой объем всей продукции (предприятия 2 и 3). Числитель средней величины $\sum_{j=1}^{4} x_j f_j$ — это объем выпуска предметов потребления всеми предприятиями — величина, которая должна сохраняться неизменной при замене разных четырех долей на среднюю долю. Расчет по данным табл. 5.3 проведен на основе известных индивидуальных значений осредняемого признака и весов.

Но исходная информация может иметь другую форму: индивидуальные значения осредняемого признака могут быть неизвестны, зато известны индивидуальные или суммарные значения объемных признаков как числителя, так и знаменателя относительной величины. Например, известно, что в акционерном сельхозяйственном предприятии было посажено 145 га картофеля и собрано с них 2595,5 т продукции. При этом совершенно не известно, сколько было собрано с каждого из 145 га в отдельности, хотя индивидуальные величины продукции, полученные на каждом гектаре, существовали объективно. Однако никакой потребности в их раздельном

Таблица 5.3

Объем и структура промышленной продукции

Номер предприятия	Объем всей продукции, млн руб., f_j	Доля товаров народного потребления, %, x_j	Объем выпуска товаров народного потребления млн руб., $x_j f_j$
1	138	75	103,5
2	650	38	247,0
3	1040	12	124,8
4	219	64	140,2
Итого	2047	30,07	615,5

учете нет; учет продукции ведется по бригадам, по отдельным полям севооборота, но не по каждому гектару. Среднюю урожайность картофеля получают делением массы собранной продукции на площадь посадки, т.е. как относительную величину, характеризующую хозяйство в целом:

$$\text{Средняя урожайность} = \frac{\text{Валовой сбор, т}}{\text{Площадь посадки, га}} = \frac{2595{,}5}{145} = 17{,}9 \text{ т/га}.$$

По отношению к предприятию это относительный затель. Но существуют и сами значения урожайности с каждого из 145 га, хотя и неучтенные. По отношению к ним 17,9 т с 1 га — это средняя величина. Такую форму определения средней арифметической величины, при которой остаются неизвестными индивидуальные значения осредняемого признака, следует называть *неявной формой средней*. Формула такой средней имеет вид:

$$\bar{x} = \frac{\sum y_i}{\sum z_i},$$

где $x_i = \frac{y_i}{z_i}$.

Свойства средней арифметической величины

Знание некоторых математических свойств средней арифметической полезно как при ее использовании, так и при ее расчете.

1. Сумма отклонений индивидуальных значений признака от его среднего значения равна нулю.

Доказательство:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) =$$

$$= x_1 + x_2 + \ldots + x_n - n\bar{x} = \sum_{i=1}^{n} x_i - n\frac{\sum_{i=1}^{n} x_i}{n} = 0.$$

Примечание. Для взвешенной средней сумма взвешенных отклонений равна нулю. Попробуйте доказать это самостоятельно.

2. Если каждое индивидуальное значение признака умножить или разделить на постоянное число, то и средняя увеличится или уменьшится во столько же раз.

Доказательство:

$$\frac{\sum_{i=1}^{n}(x_i : c)}{n} = \frac{\frac{x_1}{c} + \frac{x_2}{c} + \ldots + \frac{x_n}{c}}{n} = \frac{\frac{x_1 + x_2 + \ldots + x_n}{c}}{n} =$$

$$= \frac{x_1 + x_2 + \ldots + x_n}{n} : c = \bar{x} : c.$$

Вследствие этого свойства индивидуальные значения признака можно сократить в c раз, произвести расчет средней и результат умножить на c.

3. Если к каждому индивидуальному значению признака прибавить или из каждого значения вычесть постоянное число, то средняя величина возрастет или уменьшится на это же число.

Доказательство:

$$\frac{\sum_{i=1}^{n}(x_i + c)}{n} = \frac{(x_1 + c) + (x_2 + c) + \ldots + (x_n + c)}{n} = \frac{\sum_{i=1}^{n} x_i + nc}{n} = \bar{x} + c.$$

Это свойство полезно использовать при расчете средней величины из многозначных и слабоварьирующих значений признака, например роста группы лиц: $x_1 = 179$ см; $x_2 = 183$ см; $x_3 = 171$ см; $x_4 = 180$ см; $x_5 = 169$ см. Для вычисления среднего

роста из каждого значения вычитаем 170 см и находим среднюю из остатков: (9 + 13 + 1 + 10 − 1) : 5 = 6,4. Средний рост = 6,4 + 170 = 176,4 см.

4. Если веса средней взвешенной умножить или разделить на постоянное число, средняя величина не изменится.

Доказательство:

$$\frac{\sum\limits_{j=1}^{k} x_j \frac{f_j}{c}}{\sum\limits_{j=1}^{k} \frac{f_j}{c}} = \frac{\left(\sum\limits_{j=1}^{k} x_j f_j\right):c}{\left(\sum\limits_{j=1}^{k} f_j\right):c} = \bar{x}.$$

Используя это свойство, при расчетах следует сокращать веса на их общий сомножитель либо выражать многозначные числа весов в более крупных единицах измерения.

В табл. 5.4 приведен пример комплексного использования свойств средней арифметической для облегчения расчетов.

Средний надой молока на корову находим так:

$$\bar{x} = \left[\sum_{j=1}^{5} \frac{x_j' - 4000}{100} \cdot f_j\right] : 307 \cdot 100 + 4000 = \frac{-156}{307} \cdot 100 + 4000 = 3949 \text{ кг}.$$

5. Сумма квадратов отклонений индивидуальных значений признака от средней арифметической меньше, чем от любого другого числа.

Таблица 5.4

Расчет средней продуктивности коров на ферме

Группы коров по надою за год, кг, x_j	Число коров, f_j	Середина интервала, кг, x_j'	$\frac{x_j' - 4000}{100}$	$\frac{x_j' - 4000}{100} \cdot f_j$
3000—3400	43	3200	−8	−344
3400—3800	71	3600	−4	−284
3800—4200	102	4000	0	0
4200—4600	64	4400	4	256
4600—5000	27	4800	8	216
Итого	307	—	—	−156

Доказательство:

Составим сумму квадратов отклонений от переменной a:

$$f(a) = \sum_{i=1}^{n} (x_i - a)^2.$$

Для того чтобы найти экстремум этой функции, нужно ее производную по a приравнять нулю:

$$\frac{df}{da} = 2\sum_{i=1}^{n} (x_i - a)(-1) = 0.$$

Отсюда имеем:

$$\sum_{i=1}^{n} (x_i - a) = 0; \quad a\sum_{i=1}^{n} (1) - \sum_{i=1}^{n} (x_i) = 0;$$

$$na = \sum_{i=1}^{n} x_i; \quad a = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \bar{x}.$$

Таким образом, экстремум суммы квадратов отклонений достигается при $a = \bar{x}$. Поскольку логически ясно, что максимума функция не может иметь, этот экстремум является минимумом.

Применение простой и взвешенной средних

Простая и взвешенная средние величины различаются не только по величине (не всегда), по способу вычисления, но и по своей роли в решении различных задач статистического анализа. Рассмотрим, например, среднюю величину урожайности картофеля в группе хозяйств. Если эта средняя при решении поставленной задачи входит в систему показателей площади посадки, валового сбора, себестоимости, суммы затрат и других характеристик производства, то следует применять взвешенную среднюю, так как произведение невзвешенной средней на общую сумму площадей не даст суммы валового сбора.

Если же нас интересуют такие задачи, как измерение вариации урожайности между хозяйствами или связь урожайности с дозой органических удобрений, то следует применять простую среднюю величину урожайности, полностью абстрагируясь от размеров площадей посадки. Иначе на полученный результат повлияют различия площадей, совершенно не касающиеся этого признака. Точно так же, если необходимо изучить колебания урожайности за ряд лет и выявить их связь с температурой июня и суммой осадков за лето, нужно применять простую среднюю урожайность за ряд лет, абстрагируясь от различия размеров площадей в разные годы.

Чтобы правильно применять средние величины, следует знать, *от каких причин зависит различие между простой и взвешенной средними*. Рассмотрим этот вопрос на примере средней арифметической. Пусть $\bar{x}$ — простая средняя, $\bar{x}_z$ — взвешенная средняя, в которой весами выступают значения признака z, n — число единиц совокупности. Отклонения индивидуальных значений признака x_i от простой средней $\bar{x}$ обозначим: $\Delta_{xi} = x_i - \bar{x}$. Отклонения признака-веса: $\Delta_{zi} = z_i - \bar{z}$. Тогда индивидуальные значения признаков x и z можно выразить через их средние и отклонения: $x_i = \bar{x} + \Delta_{xi}$; $z_i = \bar{z} + \Delta_{zi}$, а взвешенную среднюю $\bar{x}_z$ представить в виде

$$\bar{x}_z = \sum_{i=1}^{n} (\bar{x} + \Delta_{xi}) : (\bar{z} + \Delta_{zi}) : \sum_{i=1}^{n} z_i. \qquad (5.3)$$

Перемножим величины в скобках и просуммируем почленно, имея в виду, что $\sum_{i=1}^{n} \bar{x} = n\bar{x}$; $\sum_{i=1}^{n} \bar{z} = n\bar{z}$. Средние величины можно вынести за знак суммирования как константы. Получим:

$$\bar{x}_z = \frac{n\bar{x}\bar{z} + \bar{x}\sum_{i=1}^{n} \Delta_{zi} + \bar{z}\sum_{i=1}^{n} \Delta_{xi} + \sum_{i=1}^{n} \Delta_{xi}\Delta_{zi}}{n\bar{z}}.$$

Поскольку суммы отклонений индивидуальных значений признака от средней арифметической согласно первому ее свойству равны нулю, то второе и третье слагаемые числителя также равны нулю.

Остается:

$$\bar{x}_z = \bar{x} + \frac{\sum_{i=1}^{n} \Delta_{xi}\Delta_{zi}}{n\bar{z}}. \qquad (5.4)$$

Числитель второго слагаемого в формуле (5.4) — это числитель коэффициента корреляции между осредняемым и весовым признаками (формулы (9.11) и (9.14)). Подставив выражение коэффициента корреляции r_{xz} в формулу (5.4), получим:

$$\bar{x}_z = \bar{x} + \frac{n\sigma_x\sigma_z r_{xz}}{n\bar{z}} = \bar{x} + \sigma_x v_z r_{xz}. \qquad (5.5)$$

Итак, средняя арифметическая взвешенная равна простой средней плюс произведение среднего квадратического отклонения осредняемого признака на коэффициент вариации весового признака и на коэффициент корреляции между этими признаками. Если обе части равенства (5.5) разделить на простую среднюю $\bar{x}$, получим:

$$\frac{\bar{x}_z}{\bar{x}} = 1 + v_x v_z r_{xz}.$$

О среднем квадратическом отклонении и коэффициенте вариации см. ниже в этой главе.

Из формулы (5.5) следует, что взвешенная средняя равна простой в трех случаях:

- если не варьирует изучаемый признак, $\sigma_x = 0$ — тривиальная ситуация, когда и сами средние не нужны;
- при условии, что не варьирует признак-вес, $v_z = 0$;
- в случаях, когда между осредняемым и признаком-весом нет линейной корреляции, $r_{xz} = 0$.

Взвешенная средняя больше простой, если эта корреляция прямая. Взвешенная средняя меньше простой средней, если эта корреляция обратная.

5.3. Другие формы средних величин

Средняя квадратическая величина

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменной сумму квадратов исходных величин, то средняя будет являться *квадратической средней величиной* ($\bar{x}_{кв}$). Ее формула такова:

$$\bar{x}_{кв} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}. \qquad (5.6)$$

Например, имеются три участка земельной площади со сторонами квадрата: $x_1 = 100$ м; $x_2 = 200$ м; $x_3 = 300$ м. Заменяя разные значения длины сторон на среднюю, мы, очевидно, должны исходить из сохранения общей площади всех участков. Арифметическая средняя величина (100 + 200 + 300) : 3 = 200 м не удовлетворяет этому условию, так как общая площадь трех участков со стороной 200 м была бы равна: $3 \cdot (200 \text{ м})^2 = 120\,000 \text{ м}^2$. В то же время площадь исходных трех участков равна: $(100 \text{ м})^2 + (200 \text{ м})^2 + (300 \text{ м})^2 = 140\,000 \text{ м}^2$. Правильный ответ дает квадратическая средняя:

$$\bar{x}_{кв} = \sqrt{\frac{(100)^2 + (200)^2 + (300)^2}{3}} = 216 \text{ м}^2.$$

Во второй части главы будет показано, что главной сферой применения квадратической средней в силу свойства 5 средней арифметической величины является измерение вариации признака в совокупности.

Аналогично если по условиям задачи необходимо сохранить неизменной сумму кубов индивидуальных значений признака при их замене на среднюю величину, мы приходим к *средней кубической*, имеющей вид:

$$\bar{x}_{куб} = \sqrt[3]{\frac{\sum_{i=1}^{n} x_i^3}{n}}. \qquad (5.7)$$

Средняя геометрическая величина

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменным произведение индивидуальных величин, то следует применить *геометрическую среднюю величину*. Ее формула такова:

$$\bar{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}. \qquad (5.8)$$

Основное применение геометрическая средняя находит при определении средних темпов роста, о чем сказано в главе 12. Пусть, например, в результате инфляции за первый год цена товара возросла в 2 раза к предыдущему году, а за второй год еще в 3 раза к уровню предыдущего года. Ясно, что за два года цена выросла в 6 раз. Каков средний темп роста цены за год? Арифметическая средняя здесь непригодна, ибо если за год цены возросли бы в $\frac{2+3}{2} = 2{,}5$ раза, то за два года цена возросла бы в $2{,}5 \cdot 2{,}5 = 6{,}25$ раза, а не в 6 раз. Геометрическая средняя дает правильный ответ: $\sqrt{6} = 2{,}45$ раза.

Геометрическая средняя величина дает наиболее правильный результат осреднения, если задача состоит в нахождении такого значения признака, который качественно был бы равноудален как от максимального, так и от минимального значения признака. Например, если максимальный размер выигрыша в лотерее составляет 1 000 000 руб., а минимальный — 100 руб., то какую величину выигрыша можно считать средней? Средняя арифметическая явно непригодна, она составляет 500 050 руб., а это, как и 1 000 000 руб., крупный, никак не средний выигрыш; он качественно однороден с максимальным и резко отличен от минимального. Не дают верного ответа ни квадратическая средняя (707 107 руб.), ни кубическая (793 699 руб.), ни рассматриваемая далее гармоническая средняя (199,98 руб.), слишком близкая к минимальному значению. Только геометрическая средняя дает верный с точки зрения экономики и логики ответ: $\sqrt{100 \cdot 1\,000\,000} = 10\,000$ руб. Десять тысяч — не миллион, но и не сотня! Это действительно нечто среднее между ними.

Средняя гармоническая величина

Если по условиям задачи необходимо, чтобы при осреднении неизменной оставалась сумма величин, обратных индивидуальным значениям признака, то средняя величина является гармонической средней.

Формула средней гармонической величины такова:

$$\bar{x}_{\text{гарм}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}. \qquad (5.9)$$

Например, автомобиль с грузом от предприятия до склада ехал со скоростью 40 км/ч, а обратно порожняком со скоростью 60 км/ч. Какова средняя скорость автомобиля за обе поездки? Пусть расстояние перевозки составляло s км. Никакой роли при расчете средней скорости величина s не играет. При замене индивидуальных значений скорости $x_1 = 60$ и $x_2 = 40$ на среднюю величину необходимо, чтобы неизменной величиной осталось время, затраченное на обе поездки, иначе средняя скорость может оказаться любой — от скорости черепахи до скорости света.

Время поездок есть $s/x_1 + s/x_2$. Итак, $s/\bar{x} + s/\bar{x} = s/x_1 + s/x_2$. Сократив все члены равенства на s, получим: $1/\bar{x} + 1/\bar{x} = 1/x_1 + 1/x_2$, т.е. выполняется условие гармонической средней. Подставляя x_1 и x_2, получаем:

$$\frac{2}{\bar{x}} = \frac{1}{60} + \frac{1}{40}, \quad \bar{x} = \frac{2}{\frac{1}{60} + \frac{1}{40}} = \frac{2 \cdot 120}{5} = 48 \text{ км/ч}.$$

Арифметическая средняя 50 км/ч неверна, так как приводит к другому времени движения, чем на самом деле. Если расстояние равно 96 км, то реальное время движения составит:

$$\frac{96}{60} + \frac{96}{40} = 1{,}6 \text{ ч} + 2{,}4 \text{ ч} = 4 \text{ ч}.$$

То же время дает гармоническая средняя:

$$\frac{(96 \cdot 2)}{48} = 4 \text{ ч}.$$

Понятие степенной средней. Соотношение между формами средних величин

Все рассмотренные выше виды средних величин принадлежат к общему типу *степенных средних*. Различаются они лишь показателем. Степенная средняя степени k есть корень k-й степени из частного от деления суммы индивидуальных значений признака в k-й степени на число индивидуальных значений:

$$\bar{x} = \sqrt[k]{\frac{\sum_{i=1}^{n} x_i^k}{n}}. \qquad (5.10)$$

При $k = 1$ получаем арифметическую среднюю, при $k = 2$ — квадратическую, при $k = 3$ — кубическую, при $k = 0$ — геометрическую, при $k = -1$ — гармоническую среднюю. Чем выше показатель степени k, тем больше значение средней величины (если индивидуальные значения признака варьируют). Если все исходные значения признака равны, то и все средние равны этой константе. Итак, имеем следующее соотношение, которое называется *правилом мажорантности средних*:

$$\bar{x}_{\text{гарм}} \leqslant \bar{x}_{\text{геом}} \leqslant \bar{x}_{\text{арифм}} \leqslant \bar{x}_{\text{квадр}} \leqslant \bar{x}_{\text{куб}} \qquad (5.11)$$

или

$$\bar{x}_{-1} \leqslant \bar{x}_0 \leqslant \bar{x}_1 \leqslant \bar{x}_2 \leqslant \bar{x}_3.$$

Пользуясь этим правилом, статистика может в зависимости от настроения и желания ее «знатока» либо «утопить», либо «выручить» студента, получившего на сессии оценки 2 и 5. Каков его средний балл?

Если судить по средней арифметической, то средний балл равен 3,5. Но если декан желает «утопить» несчастного и вычислит среднюю гармоническую

$$\bar{x}_{\text{гарм}} = \frac{2}{\frac{1}{2} + \frac{1}{5}} = \frac{20}{7} = 2{,}86,$$

то студент остается в среднем двоечником, не дотянувшим до тройки. Однако студенческий комитет может возразить декану и представить среднюю кубическую величину:

$$\bar{x}_{\text{куб}} = \sqrt[3]{\frac{2^3 + 5^3}{2}} = \sqrt[3]{66{,}5} = 4{,}05.$$

Студент уже выглядит «хорошистом» и даже претендует на стипендию! И только в том случае, если лентяй провалил оба экзамена, статистика помочь не в состоянии: увы, все средние из двух двоек равны все той же двойке!

5.4. Средняя величина как выражение закономерности

После того как мы познакомились с различными видами и формами средних величин, включая и неявную их форму, можно перейти к понятию о средних. В широком понимании термина средней величиной является всякий обобщающий показатель, характеризующий обобщенное значение признака, связи признаков, их динамики и структуры в совокупности массовых явлений.

Так, средними в широком смысле слова являются такие показатели, как доля мужчин в общем числе жителей страны (ведь эта доля разная в разных регионах), плотность населения, коэффициент смертности, ожидаемая продолжительность жизни родившихся в данном году и др. Рассматриваемые далее в этой главе показатели вариации признака в совокупности, а также в гл. 9 показатели корреляционной связи тоже средние в широком смысле слова, так как измеряют среднее различие между значениями одного признака у разных единиц совокупности, или среднюю связь вариации одного признака с вариацией другого.

В такой же степени средними являются и показатели темпов роста продукции промышленности, или национального дохода страны, обобщающие темпы разных отраслей и регионов; средними являются меры колеблемости урожайности за ряд лет (гл. 12), обобщающие влияние на урожайность разных лет метеорологических и экономических условий производства.

Понятие средней величины в широком смысле слова сближается с такой философской категорией, как закон («закон есть общее в явлениях»), закономерность. Это далеко не случайное родство. Рассмотрим сущность процесса осреднения на примере арифметической средней согласно формуле (5.1). Среднюю считаем типической, определенной по однородной совокупности. Однородность индивидуальных значений признака — это проявление их общих свойств, обусловленных основными условиями и закономерностями массового процесса, порождающего данную совокупность. Однако, кроме общих условий и закономерности, на каждую единицу совокупности влияют индивидуальные, особенные условия, случайные события, не связанные причинно с общей закономерностью. Поэтому индивидуальные значения признака x_i можно представить как состоящие из элемента, обусловленного общей закономерностью для всех единиц совокупности (обозначим этот элемент c), так и элемента Δ_i, индивидуального для каждой единицы совокупности. Итак, $x_i = c + \Delta_i$, где Δ_i может быть как положительной, так и отрицательной величиной, как малой, так и большой величиной в сравнении с c.

Теперь вычислим среднее значение признака для совокупности из n единиц:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} (c + \Delta_i)}{n} = \frac{nc + \sum_{i=1}^{n} \Delta_i}{n} = c + \bar{\Delta}.$$

Итак, средняя величина признака слагается из элемента, выражающего закономерность, общую для всей совокупности, и из средней величины элементов, отражающих индивидуальные условия отдельных единиц этой совокупности. Элементы Δ_i могут иметь положительные и отрицательные, большие и малые значения. При осреднении они согласно закону больших чисел взаимопогашаются в зависимости от объема совокупности — тем в большей мере, чем больше объем совокупности n. Об этом говорит формулировка закона больших чисел, данная великим русским математиком П. Л. Чебышевым (1821—1894). Чем больше объем однородной совокупности, тем полнее взаимопогашение случайных (по отношению к совокупности в целом и ее законам) элементов при-

знака x; полнее и надежнее, с большей вероятностью среднее значение признака измеряет действие общих для совокупности закономерностей.

Однако случайная вариация индивидуальных величин признаков — это не только некоторая помеха, туман, «шум» в информационном смысле, затрудняющие познание закономерности. Вариация — неотъемлемая, необходимая черта, свойство массовых явлений, имеющая громадное самостоятельное значение в развитии природы и общества.

Создатель учения о средних величинах бельгийский статистик А. Кетле по этому поводу писал следующее: «В мире существует общий закон, предназначенный как бы для того, чтобы разливать жизнь во Вселенной; в силу этого закона все живущее подлежит бесконечному разнообразию... Каждый предмет подвержен флюктуациям»[1].

В следующих подразделах данной главы переходим к методам статистического изучения этого «общего закона Вселенной» — вариации массовых явлений и их признаков.

5.5. Вариация массовых явлений

Вариацией значений какого-либо признака в совокупности называется различие его значений у разных единиц данной совокупности в один и тот же период или момент времени.

В отличие от вариации различия значений признака у одного и того же объекта, у одной и той же единицы совокупности в разные периоды или моменты времени следует называть *изменениями во времени и колебаниями*. Методы их измерения и изучения отличаются принципиально от методов измерения вариации (гл. 10).

Причиной вариации являются разные условия существования разных единиц совокупности. Даже однояйцевые близнецы в процессе развития приобретают различия в росте, весе, не говоря уже о таких признаках, как специальность, образование, заработная плата (доход), число детей и т.д. Еще больше причин влияют на различия промышленных предприятий, магазинов и пр.

[1] *Кетле А.* Социальная система и законы, ею управляющие: Пер. с фр. — СПб., 1866. — С. 16.

Вариация присуща всем без исключения явлениям природы и общества, кроме законодательно закрепленных нормативных значений отдельных социальных признаков: не варьирует признак «число председателей правления АОЗТ» — все они имеют по одному председателю. Неварьирующие признаки не представляют интереса для статистики; *предметом изучения статистики является вариация.* Большинство методов статистики — это либо методы измерения вариации, либо методы абстрагирования от нее.

Вариация, несомненно, — необходимое условие существования и развития массовых явлений. Например, вариация геномов (набора генов) родительских организмов растений и животных обеспечивает жизнеспособность потомства. Близкородственный брак, т.е. слишком малая вариация геномов родителей, ведет к неполноценному потомству. Перекрестное опыление для многих растений — обязательное условие плодоношения. Гибридизация, т.е. получение потомства от неродственных, со значительной вариацией свойств сортов сельскохозяйственных растений и пород животных, — важный прием повышения урожайности и продуктивности скота.

В то же время известно, что нельзя получить потомство от организмов со слишком разными свойствами — разных видов, родов и семейств, например от кошки и собаки. Чрезмерная вариация генотипов препятствует развитию. И в промышленном производстве, особенно массовом, вариация размеров, свойств деталей, из которых собирается станок, автомашина, телевизор, должна быть введена в жесткие рамки «допусков», т.е. пренебрежимо малых величин, чтобы сборка была возможной и не страдало качество собранного агрегата.

Можно сказать, что в жизни общества, как и в природе, каждой массовой совокупности, массовому процессу присуща некоторая специфическая мера вариации ее элементов, при которой данный процесс протекает оптимально.

Для того чтобы руководитель предприятия, менеджер, научный работник могли управлять вариацией и изучать ее, статистикой разработаны специальные методы исследования, система показателей, с помощью которой вариация измеряется, характеризуются ее свойства.

5.6. Построение вариационного ряда. Виды рядов. Ранжирование данных

Первым этапом статистического изучения вариации являются построение *вариационного ряда* — упорядоченного распределения единиц совокупности по возрастающим (чаще) или по убывающим (реже) значениям признака и подсчет числа единиц с тем или иным значением признака.

Существуют три формы вариационного ряда: ранжированный, дискретный, интервальный. Вариационный ряд часто называют *рядом распределения*. Этот термин употребляется при изучении вариации как количественных, так и неколичественных признаков. Ряд распределения представляет собой *структурную группировку* (гл. 6).

Ранжированный ряд — это перечень отдельных единиц совокупности в порядке возрастания (убывания) изучаемого признака.

Ниже приведены сведения о крупных банках Санкт-Петербурга, ранжированных по размерам собственного капитала на 01.10.1999 г.

Название банка	Собственный капитал, млн руб.
Балтонэксим банк	169
Банк «Санкт-Петербург»	237
Петровский	268
Балтийский	290
Промстройбанк	1007

Если численность единиц совокупности достаточно велика, ранжированный ряд становится громоздким, а его построение, даже с помощью компьютера, занимает длительное время. В таких случаях вариационный ряд строится с помощью группировки единиц совокупности по значениям изучаемого признака.

Если признак принимает небольшое число значений, строится дискретный вариационный ряд. Примером такого ряда является распределение футбольных матчей по числу забитых мячей (см. табл. 5.1). *Дискретный вариационный ряд* — это таблица, состоящая из двух строк или граф: конкретных значений варьирующего признака x_i и числа единиц совокупности с данным значением признака f_i — частот (f — начальная буква англ. слова frequency).

Определение числа групп

Число групп в дискретном вариационном ряду определяется числом реально существующих значений варьирующего признака. Если признак принимает дискретные значения, но их число очень велико (например, поголовье скота на 1 января года в разных сельскохозяйственных предприятиях может составить от нуля до десятков тысяч голов), то строится интервальный вариационный ряд. Интервальный вариационный ряд строится и для изучения признаков, которые могут принимать любые, как целые, так и дробные значения в области своего существования. Таковы, например, рентабельность реализованной продукции, себестоимость единицы продукции, доход на одного жителя города, доля лиц с высшим образованием среди населения разных территорий и вообще все вторичные признаки, значения которых рассчитываются путем деления величины одного первичного признака на величину другого (см. гл. 3).

Интервальный вариационный ряд представляет собой таблицу, состоящую из двух граф (или строк) — интервалов признака, вариация которого изучается, и числа единиц совокупности, попадающих в данный интервал (частот), или долей этого числа от общей численности совокупности (частостей).

Наиболее часто используются два вида интервальных вариационных рядов: равноинтервальный и равночастотный. Равноинтервальный ряд применяется, если вариация признака не очень сильна, т.е. для однородной совокупности, распределение которой по данному признаку близко к нормальному закону. (Такой ряд представлен в подразд. 5.6.) Равночастотный ряд применяется, если вариация признака очень сильна, однако распределение не является нормальным, а, например, гиперболическим (табл. 5.6).

При построении равноинтервального ряда число групп выбирается так, чтобы в достаточной мере отразились разнообразие значений признака в совокупности и в то же время закономерность распределения, его форма не искажалась случайными колебаниями частот. Если групп будет слишком мало, не проявится закономерность вариации; если групп будет чрезмерно много, случайные скачки частот исказят форму распределения.

Чаще всего число групп в вариационном ряду устанавливают, придерживаясь формулы американского статистика Стерджесса

$$k \approx 1 + 3{,}32 \cdot \lg n = 1 + 1{,}44 \ln n,$$

где k — число групп;
n — численность совокупности.

Эта формула показывает, что число групп — функция объема данных.

Предположим, необходимо построить вариационный ряд распределения предприятий области по урожайности зерновых культур за какой-то год. Число сельскохозяйственных предприятий, имевших посевы зерновых культур, составило 143; наименьшее значение урожайности равно 10,7 ц/га, наибольшее — 53,1 ц/га. Имеем:

$$k \approx 1 + 3{,}32 \cdot \lg 143 = 8{,}16.$$

Поскольку число групп целое, рекомендуется построить 8 или 9 групп.

Определение величины интервала

Зная число групп, рассчитывают величину интервала:

$$i = \frac{x_{max} - x_{min}}{k}.$$

В нашем примере величина интервала составляет:
а) при 8 группах

$$i = \frac{53{,}1 - 10{,}7}{8} = 5{,}3 \text{ ц/га},$$

б) при 9 группах

$$i = \frac{53{,}1 - 10{,}7}{9} = 4{,}7 \text{ ц/га}.$$

Для построения ряда и анализа вариации значительно лучше иметь по возможности округленные значения величины интервала и его границ. Поэтому наилучшим решением будет построение вариационного ряда с 9 группами с интервалом, равным 5 ц/га. Этот вариационный ряд приведен в табл. 5.6, а его графическое изображение — на рис. 5.1.

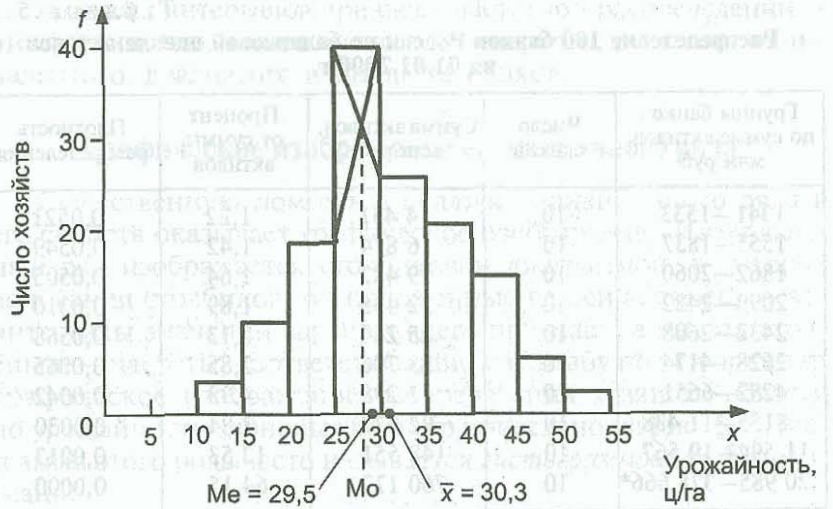


Рис. 5.1. Распределение хозяйств по урожайности

Границы интервалов могут указываться разным образом: верхняя граница предыдущего интервала повторяет нижнюю границу следующего, как показано в табл. 5.5, или не повторяет.

В последнем случае второй интервал будет обозначен как 15,1—20, третий — как 20,1—25 и т.д., т.е. предполагается, что все значения урожайности обязательно округлены до одной десятой. Кроме того, возникает нежелательное осложнение с серединой интервала 15,1—20, которая, строго говоря, уже будет равна не 17,5, а 17,55; соответственно при замене округленного интервала 40—60 на 40,1—60 вместо округленного значения его середины 50 получим 50,5. Поэтому предпочтительнее оставить интервалы с повторяющейся округленной границей и договориться, что единицы совокупности, имеющие значение признака, равное границе интервала, включаются в тот интервал, где это точное значение впервые указывается. Так, хозяйство, имеющее урожайность, равную 15 ц/га, включается в первую группу, значение 20 ц/га — во вторую и т.д.

Равночастотный вариационный ряд необходим при очень сильной вариации признака потому, что при равноинтервальном распределении большая часть единиц совокупности ока-

Таблица 5.5

Распределение 100 банков России по балансовой оценке активов на 01.01.2000 г.

Группы банков по сумме активов, млн руб.	Число банков	Сумма активов, млн руб.	Процент от суммы активов	Плотность распределения
1341—1533	10	14 431	1,22	0,0521
1555—1837	10	16 874	1,42	0,0549
1862—2060	10	19 482	1,64	0,0505
2099—2422	10	22 435	1,89	0,0310
2432—2608	10	25 244	2,13	0,0568
2628—4174	10	33 796	2,85	0,0065
4282—6651	10	51 298	4,33	0,0042
8155—11 438	10	92 872	7,84	0,0030
11 594—19 567	10	148 551	12,53	0,0013
20 985—378 666*	10	760 172	64,15	0,0000
Итого	100	1 185 158	100	—

* Сбербанк России.

жется в крайней группе (первой или последней). Часть групп не будет включать ни одной единицы совокупности. Число групп равночастотного ряда часто принимают равным 10, т.е. в каждой группе содержится 10% единиц совокупности. Такие группы называют *децильными*. Так, например, группируют домохозяйства по данным выборочного бюджетного обследования доходов населения России и субъектов федерации органы Госкомстата. Децильные группы применяются также для изучения распределений банков, промышленных компаний по размерам их активов или собственного капитала.

Распределение банков по децильным группам наглядно показывает сильную неравномерность их капиталов: 10% крупнейших банков имеют более 64% всех капиталов 100 банков. Если бы был построен равноинтервальный ряд, например, с семью группами, то ширина интервала составила бы 53 904 млн руб. и в первую группу попало бы 98 банков, один — во вторую и один — в седьмую. Ясно, что никакого смысла такой ряд не имеет.

Границы интервалов при равночастотном распределении — это фактические величины активов первого, десятого, одиннадцатого, двадцатого и так далее банков.

Графическое изображение вариационного ряда

Существенную помощь в анализе вариационного ряда и его свойств оказывает графическое изображение. Интервальный ряд изображается столбиковой диаграммой, в которой основания столбиков, расположенные на оси абсцисс, — это интервалы значений варьирующего признака, а высота столбиков — частоты, соответствующие масштабу по оси ординат. Графическое изображение распределения хозяйств области по урожайности зерновых культур приведено на рис. 5.1. Диаграмма этого рода часто называется *гистограммой* (гр. histos — ткань).

Данные табл. 5.6 и рис. 5.1 показывают характерную для многих признаков форму распределения: чаще встречаются значения средних интервалов признака, реже — крайние, малые и большие значения признака. Форма этого распределения близка к рассматриваемому в курсе математической статистики закону нормального распределения. Великий русский математик А. М. Ляпунов (1857—1918) доказал, что нор-

Таблица 5.6

Распределение хозяйств области по урожайности зерновых культур

Урожайность, ц/га, x_j	Число хозяйств, f_j	Середина интервала, ц/га, x_j'	$x_j' f_j$	Накопленная частота, f_j'
1	2	3	4	5
10—15	6	12,5	75,0	6
15—20	9	17,5	157,5	15
20—25	20	22,5	450,0	35
25—30	41	27,5	1127,5	76
30—35	26	32,5	845,0	102
35—40	21	37,5	787,5	123
40—45	14	42,5	595,0	137
45—50	5	47,5	237,5	142
50—55	1	52,5	52,5	143
Итого	143		4327,5	

мальное распределение образуется, если на варьирующую переменную влияет большое число факторов, ни один из которых не имеет преобладающего влияния. Случайное сочетание множества примерно равных факторов, влияющих на вариации урожайности зерновых культур, как природных, так и агротехнических, экономических, создает близкое к нормальному закону распределения распределение хозяйств области по урожайности.

Если имеется дискретный вариационный ряд или используются середины интервалов, то графическое изображение такого вариационного ряда называется *полигоном* (гр. polygónos < poly много + gonia угол). Каждый из вас легко построит этот график, соединяя прямыми точки с координатами x_i и f_i.

Отношение высоты полигона или гистограммы к их основанию рекомендуется в пропорции примерно 5 : 8.

Понятие частости

Если, используя данные табл. 5.6, число хозяйств с тем или иным уровнем урожайности выразить в процентах к итогу, принимая все число хозяйств (143) за 100%, то средняя урожайность может быть вычислена так:

$$\bar{x} = \sum_{j=1}^{k} x_j' w_j, \qquad (5.12)$$

где w_j — частость j-й категории вариационного ряда;

$$\sum_{j=1}^{k} w_j = 1 \text{ (или 100\%)}.$$

Таким образом, частость — это относительное выражение частоты.

Кумулятивное распределение

Преобразованной формой вариационного ряда является *ряд накопленных частот* (табл. 5.6, графа 5). Это ряд значений числа единиц совокупности с меньшими и равными нижней границе соответствующего интервала значениями признака.

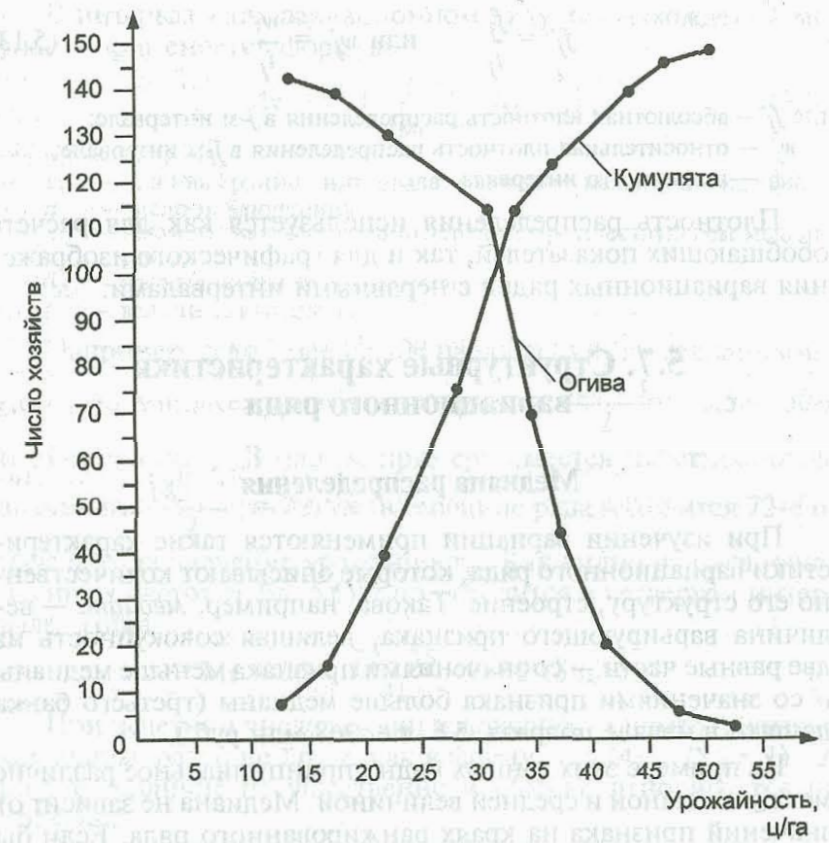


Рис. 5.2. Кумулята и огива распределения хозяйств по урожайности

Такой ряд называется *кумулятивным*. Можно построить кумулятивное распределение «не меньше, чем», а можно «больше, чем». В первом случае график кумулятивного распределения называется кумулятой, во втором — огивой (рис. 5.2).

Плотность распределения

Если приходится иметь дело с вариационным рядом с неравными интервалами, то для сопоставимости нужно частоты, или частости, привести к единице интервала. Полученное отношение называется плотностью распределения.

$$f_j' = \frac{f_j}{i_j} \quad \text{или} \quad w_j' = \frac{w_j}{i_j}, \qquad (5.13)$$

где f_j' — абсолютная плотность распределения в j-м интервале;
w_j' — относительная плотность распределения в j-м интервале;
i_j — ширина j-го интервала.

Плотность распределения используется как для расчета обобщающих показателей, так и для графического изображения вариационных рядов с неравными интервалами.

5.7. Структурные характеристики вариационного ряда

Медиана распределения

При изучении вариации применяются такие характеристики вариационного ряда, которые описывают количественно его структуру, строение. Такова, например, *медиана* — величина варьирующего признака, делящая совокупность на две равные части — со значениями признака меньше медианы и со значениями признака больше медианы (третьего банка из пяти в начале подразд. 5.6, т.е. 268 млн руб.).

На примере этих данных видно принципиальное различие между медианой и средней величиной. Медиана не зависит от значений признака на краях ранжированного ряда. Если бы капитал крупнейшего банка Санкт-Петербурга был в десять раз больше, величина медианы не изменилась бы. Поэтому часто медиану используют как более надежный показатель типичного значения признака, нежели арифметическая средняя, если ряд значений неоднороден, включает резкие отклонения от средней. В данном ряду средняя величина собственного капитала равна 394 млн руб., сложилась под влиянием наибольшей варианты. 80% банков имеют капитал меньше среднего и лишь 20% — больше. Вряд ли такую среднюю можно считать типичной величиной. При четном числе единиц совокупности за медиану принимают арифметическую среднюю величину из двух центральных вариант, например при 10 значениях признака — среднюю из пятого и шестого значений в ранжированном ряду.

В интервальном вариационном ряду для нахождения медианы применяется формула

$$\text{Me} = x_e + \frac{i}{f_{\text{Me}}}\left(\frac{n}{2} - f_{\text{Me}-1}\right), \qquad (5.14)$$

где Me — медиана;
x_e — нижняя граница интервала, в котором находится медиана;
n — число наблюдений;
$f_{\text{Me}-1}$ — накопленная частота в интервале, предшествующем медианному;
f_{Me} — частота в медианном интервале;
i — величина интервала.

Например, если имеется 100 наблюдений, то медианными, т.е. стоящими в середине ряда, являются: $\frac{100+1}{2} = 50{,}5$ — 50-я и 51-я единицы. В нашем примере имеется нечетное число значений: $\frac{143+1}{2} = 72$, т.е. в середине ряда находится 72-е от начала ряда значение урожайности. Как видно из ряда накопленных частот (табл. 5.6), оно находится в четвертом интервале. Тогда

$$\text{Me} = 25 + \frac{72-35}{41} \cdot 5 = 29{,}5 \text{ ц/га}.$$

При нечетном числе единиц совокупности номер медианы, как видим, равен не $\Sigma f_j : 2$, как в формуле (5.14), а $(\Sigma f_j + 1) : 2$, но это различие несущественно и обычно игнорируется на практике.

В равноинтервальном ряду медиана — это середина среднего интервала при их нечетном числе или средняя арифметическая из границ двух средних интервалов при их четном числе.

В дискретном вариационном ряду медианой следует считать значение признака в той группе, в которой накопленная частота превышает половину численности совокупности. Например, для данных табл. 5.1 медианой числа забитых за игру мячей будет два.

Квартили распределения

Аналогично медиане вычисляются значения признака, делящие совокупность на четыре равные по числу единиц части. Эти величины называются *квартилями* и обозначаются за-

главной латинской буквой Q с подписным значком номера квартиля. Ясно, что Q_2 совпадает с Me. Для первого и третьего квартилей приводим формулы и расчет по данным табл. 5.6.

$$Q_1 = x_0 + \frac{\left(\sum_{j=1}^{k} f_j/4\right) - f'_{Q_1-1}}{f_{Q_1}} i = 25 + \frac{(35{,}75 - 35)}{41} \cdot 5 = 25{,}09 \text{ ц/га};$$

$$Q_3 = x_0 + \frac{\left(3\sum_{j=1}^{k} f_j/4\right) - f'_{Q_3-1}}{f_{Q_3}} i = 35 + \frac{(107{,}25 - 102)}{21} \cdot 5 = 36{,}25 \text{ ц/га}.$$

Поскольку Q_2 = Me = 29,5 ц/га, видно, что различие между первым квартилем и медианой меньше, чем между медианой и третьим квартилем. Этот факт свидетельствует о наличии некоторой несимметричности в средней области распределения, что заметно и на рис. 5.1.

Значения признака, делящие ряд на пять равных частей, называют *квинтилями*, на десять частей — *децилями*, на сто частей — *перцентилями*. Поскольку эти характеристики применяются лишь при необходимости подробного изучения структуры вариационного ряда, приводить их формулы и расчет не будем.

Мода распределения

Бесспорно, важное значение имеет такая величина признака, которая встречается в изучаемом ряду, в совокупности чаще всего. Такую величину принято называть *модой* и обозначать Mo. В дискретном ряду мода определяется без вычисления как значение признака с наибольшей частотой. Например, по данным табл. 5.1 чаще всего за футбольный матч было забито два мяча — 53 раза. Модой является число два. Обычно встречаются ряды с одним модальным значением признака. Если два или несколько равных (и даже несколько различных, но больших, чем соседние) значений признака имеются в вариационном ряду, он считается соответственно бимодальным («верблюдообразным») либо мультимодальным. Это говорит

о неоднородности совокупности, возможно, представляющей собой агрегат нескольких совокупностей с разными модами.

Так и в толпе туристов, приехавших из разных стран, вместо одной, преобладающей среди местных жителей модной одежды можно встретить смесь «мод», принятых у разных народов мира.

В интервальном вариационном ряду, тем более при непрерывной вариации признака, строго говоря, каждое значение признака встречается только один раз. Модальным интервалом является интервал с наибольшей частотой. Внутри этого интервала находят условное значение признака, вблизи которого плотность распределения, т.е. число единиц совокупности, приходящееся на единицу измерения варьирующего признака, достигает максимума. Это условное значение и считается *точечной модой*. Логично предположить, что такая точечная мода располагается ближе к той из границ интервала, за которой частота в соседнем интервале больше частоты в интервале за другой границей модального интервала. Отсюда имеем обычно применяемую формулу

$$\text{Mo} = x_0 + \frac{f_{\text{Mo}} - f_{\text{Mo}-1}}{(f_{\text{Mo}} - f_{\text{Mo}-1}) + (f_{\text{Mo}} - f_{\text{Mo}+1})} \cdot i, \qquad (5.15)$$

где x_0 — нижняя граница модального интервала;
f_{Mo} — частота в модальном интервале;
$f_{\text{Mo}-1}$ — частота в предыдущем интервале;
$f_{\text{Mo}+1}$ — частота в следующем интервале за модальным;
i — величина интервала.

По данным табл. 5.6 рассчитаем моду:

$$\text{Mo} = 25 + \frac{(41 - 20)}{(41 - 20) + (41 - 26)} \cdot 5 = 27{,}9 \text{ ц/га}.$$

Вычисление моды в интервальном ряду весьма условно. Приближенно Mo может быть определена графически (см. рис. 5.1).

В равноинтервальном ряду при расчете моды (5.5) следует использовать плотность распределения.

К изучению структуры вариационного ряда средняя арифметическая величина тоже имеет отношение, хотя основное значение этого обобщающего показателя другое. В ряду распределения хозяйств по урожайности (табл. 5.6) средняя ве-

личина урожайности вычисляется как взвешенная по частоте середина интервалов x' (по формуле (5.2)):

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j' f_j}{\sum_{j=1}^{k} f_j} = \frac{4327{,}5}{143} = 30{,}3 \text{ ц/га}.$$

Соотношение между средней величиной, медианой и модой

Различие между средней арифметической величиной, медианой и модой в данном распределении невелико. Если распределение по форме близко к нормальному закону, то медиана находится между модой и средней величиной, причем ближе к средней, чем к моде.

При правосторонней асимметрии $\bar{x} > \text{Me} > \text{Mo}$;
при левосторонней асимметрии $\bar{x} < \text{Me} < \text{Mo}$.

Для умеренно асимметричных распределений справедливо равенство: $|\text{Mo} - \bar{x}| = 3|\text{Me} - \bar{x}|$.

5.8. Показатели размера и интенсивности вариации

Абсолютные средние размеры вариации

Следующим этапом изучения вариации признака в совокупности является измерение характеристик силы, величины вариации. Простейшим из них может служить *размах,* или *амплитуда вариации,* — абсолютная разность между максимальным и минимальным значениями признака из имеющихся в изучаемой совокупности значений. Таким образом, размах вариации вычисляется по формуле

$$R = x_{\max} - x_{\min}. \qquad (5.16)$$

Поскольку величина размаха характеризует лишь максимальное различие значений признака, она не может измерять закономерную силу его вариации во всей совокупности. Предназначенный для данной цели показатель должен учитывать и обобщать все различия значений признака в совокупности без исключения. Число таких различий равно числу

сочетаний по два из всех единиц совокупности, по данным табл. 5.6 оно составит: $C_{143}^2 = 10\,153$. Однако нет необходимости рассматривать, вычислять и осреднять все отклонения. Проще использовать среднюю из отклонений отдельных значений признака от среднего арифметического значения признака, а таковых всего 143. Но среднее отклонение значений признака от средней арифметической величины согласно известному свойству последней равно нулю. Поэтому показателем силы вариации выступает не алгебраическая средняя отклонений, а средний модуль отклонения, или *среднее линейное отклонение.*

Этот показатель рассчитывается по формуле
для несгруппированных данных

$$a = \sum_{i=1}^{n} |x_i - \bar{x}| : n. \qquad (5.17)$$

для сгруппированных данных

$$a = \frac{\sum_{j=1}^{k} |x_j' - \bar{x}| f_j}{\sum_{j=1}^{k} f_j}, \qquad (5.18)$$

где x_j' — середина j-го интервала переменной x;
$\bar{x}$ — среднее значение переменной x;
f_j — частота j-го интервала;
k — число групп.

По данным табл. 5.6 $a = \frac{980{,}2}{143} = 6{,}85$ ц/га.

Это означает, что в среднем урожайность в изучаемой совокупности хозяйств отклонялась от средней урожайности по области на 6,85 ц/га. Простота расчета и интерпретации составляют положительные стороны данного показателя, однако математические свойства модулей «плохие»: их нельзя поставить в соответствие с каким-либо вероятностным законом, в том числе и с нормальным распределением, параметром которого является не средний модуль отклонений, а *среднее квадратическое отклонение* (в англоязычных программах для ПЭВМ называемое «The standard deviation», сокращенно s.d.

или просто s, в русскоязычных — СКО). В статистической литературе среднее квадратическое отклонение от средней величины принято обозначать малой (строчной) греческой буквой сигма (σ) или *s* (гл. 7):

для несгруппированных данных

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}, \qquad (5.19)$$

для сгруппированных данных

$$\sigma = \sqrt{\frac{\sum_{j=1}^{k}(x_j' - \bar{x})^2 f_j}{\sum_{j=1}^{k} f_j}}. \qquad (5.20)$$

По данным табл. 5.6 среднее квадратическое отклонение урожайности зерновых составило:

$$\sigma = \sqrt{\frac{10185}{143}} = 8{,}44 \text{ ц/га}.$$

Следует указать, что некоторое округление средней величины и середин интервалов, например до целых, мало отражается на величине σ, которая составила бы при этом 8,55 ц/га.

Среднее квадратическое отклонение по величине в реальных совокупностях всегда больше среднего модуля отклонений. Соотношение σ : *a* зависит от наличия в совокупности резких, выделяющихся отклонений и может служить индикатором «засоренности» совокупности неоднородными элементами: чем это соотношение больше, тем сильнее подобная «засоренность». Для нормального закона распределения σ : *a* ≈ 1,2.

Понятие дисперсии

Квадрат среднего квадратического отклонения дает величину *дисперсии* σ^2. Формула дисперсии:

для несгруппированных данных

$$\sigma^2 = \frac{\sum_{(i)}(x_i - \bar{x})^2}{n} \qquad (5.21)$$

или

$$\sigma^2 = \overline{x^2} - \bar{x}^2, \qquad (5.22)$$

для сгруппированных данных

$$\sigma^2 = \frac{\sum_{(j)}(x_j' - \bar{x})^2 f_j}{\sum_{(j)} f_j} \qquad (5.23)$$

или

$$\sigma^2 = \overline{x^2} - \bar{x}^2. \qquad (5.24)$$

Равенство результатов расчетов по формулам (5.21) и (5.22), (5.23) и (5.24) выполняется только при точном значении средней арифметической величины. Если же средняя округлена, то дисперсию следует вычислять только по формулам:

простая

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n};$$

взвешенная

$$\sigma^2 = \frac{\sum_{j=1}^{k}(x_j' - \bar{x})^2 f_j}{\sum_{j=1}^{k} f_j}.$$

Расчет по формулам (5.21) и (5.23) приведет к погрешности дисперсии того же порядка, что и погрешность, допущенная при округлении средней величины. Математик В. С. Итенберг показал, что расчет по формулам (5.22) и (5.24) приводит к погрешности дисперсии, на порядки большей, нежели допущенная при расчете средней, что видно из приведенного ниже примера (табл. 5.7).

Таблица 5.7

Расчет дисперсии

Номер единицы совокупности	x_i	Точная		x_i^2	Округленная	
		$x_i - \bar{x}$	$(x_i - \bar{x})^2$		$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	20	−10,6	112,36	400	−11	121
2	25	−5,6	31,36	625	−6	36
3	30	−0,6	0,36	900	−1	1
4	38	7,4	54,76	1444	7	49
5	40	9,4	88,36	1600	9	81
Σ	153	0	287,2	4969	−2	288

$\bar{x}_{\text{точная}} = \frac{153}{5} = 30{,}6$; $\bar{x}_{\text{округл.}} = 31$.

При точной величине $\bar{x}$: $\sigma^2 = \frac{287{,}2}{5} = 57{,}44$; или: $\sigma^2 = \frac{4969}{5} - 30{,}6^2 = 57{,}44$.

При округленной величине $\bar{x}$: $\sigma^2 = \frac{288}{5} = 57{,}6$ — погрешность 0,3%; $\sigma^2 = \bar{x}^2 - (\bar{x})^2 = \frac{4969}{5} - 31^2 = 993{,}8 - 961 = 32{,}8$, т.е. погрешность составляет 43% величины дисперсии.

На дисперсии основаны практически все методы математической статистики. Большое практическое значение имеет правило сложения дисперсий (гл. 6).

Другие меры вариации

Еще одним показателем силы вариации, характеризующим ее не по всей совокупности, а лишь в центральной части, служит *среднее квартильное расстояние*, т.е. средняя величина разности между квартилями, обозначаемое далее как q:

$$q = \frac{(Q_3 - \text{Me}) + (\text{Me} - Q_1)}{2} = \frac{Q_3 - Q_1}{2}. \quad (5.25)$$

Для распределения сельскохозяйственных предприятий по урожайности в табл. 5.6 $q = (36{,}25 - 25{,}09) = 5{,}58$ ц/га. Сила вариации в центральной части совокупности, как правило, меньше, чем в целом по всей совокупности. Соотношение между средним модулем отклонений и средним квартильным отклонением также служит для изучения структуры вариации: большое значение такого соотношения говорит о наличии слабоварьирующего «ядра» и сильно рассеянного вокруг этого ядра окружения, или «гало» в изучаемой совокупности. Для данных табл. 5.6 соотношение $a : q = 1{,}23$, что говорит о небольшом различии силы вариации в центральной части совокупности и на ее периферии.

Для оценки интенсивности вариации и для сравнения ее в разных совокупностях и тем более для разных признаков необходимы *относительные показатели вариации*. Они вычисляются как отношения абсолютных показателей силы вариации, рассмотренных ранее, к средней арифметической величине признака. Получаем следующие показатели:

1) относительный размах вариации ρ (*коэффициент осцилляции*):

$$\rho = R : \bar{x}; \quad (5.26)$$

2) относительное отклонение по модулю, m:

$$m = a : \bar{x}; \quad (5.27)$$

3) коэффициент вариации как относительное квадратическое отклонение, v:

$$v = \sigma : \bar{x}; \quad (5.28)$$

4) относительное квартильное расстояние, d:

$$d = q : \bar{x}, \quad (5.29)$$

где q — среднее квартильное расстояние.

Для вариации урожайности по данным табл. 5.6 эти показатели составляют:

$\rho = 42{,}4 : 30{,}3 = 1{,}4$, *или* 140%;

$m = 6{,}85 : 30{,}3 = 0{,}226$, или 22,6%;

$v = 8{,}44 : 30{,}3 = 0{,}279$, или 27,9%;

$d = 5{,}58 : 30{,}3 = 0{,}184$, или 18,4%.

Оценка степени интенсивности вариации возможна только для каждого отдельного признака и совокупности определенного состава. Так, для совокупности сельскохозяйственных предприятий вариация урожайности в одном и том же природном регионе может быть оценена как слабая, если $v < 10\%$, умеренная при $10\% < v < 25\%$ и сильная при $v > 25\%$.

Напротив, вариация роста в совокупности взрослых мужчин или женщин уже при коэффициенте, равном 7%, должна быть оценена и воспринимается людьми как сильная. Таким образом, оценка интенсивности вариации состоит в сравнении наблюдаемой вариации с некоторой обычной ее интенсивностью, принимаемой за норматив. Мы привыкли к тому, что урожайность, заработок или доход на душу населения, число жилых комнат в здании могут различаться в несколько и даже десятки раз, но различие роста людей в полтора раза уже воспринимается как очень сильное.

Различная сила, интенсивность вариации обусловлены объективными причинами. Например, цена продажи доллара США в одном из коммерческих банков Санкт-Петербурга на 1 января 2003 г. варьировала от 31.87 руб./долл. до 32.13 руб./долл. при средней цене 32 руб. за доллар США. Относительный размах вариации $\rho = [32.13 - 31.87] = 26$ коп. : 32 руб. $= 0{,}8\%$. Такая малая вариация вызвана тем, что при значительном различии курса доллара немедленно произошел бы отток покупателей из «дорогого» банка в более «дешевые». Напротив, цена килограмма картофеля или говядины в разных регионах России варьирует очень сильно — на десятки процентов и более. Это объясняется разными затратами на доставку товара из региона-производителя в регион-потребитель, т.е. пословицей «Телушка за морем — полушка, да рубль перевоз».

5.9. Моменты распределения и показатели его формы

Центральные моменты распределения

Для дальнейшего изучения характера вариации используются средние значения разных степеней отклонений отдельных величин признака от его средней арифметической величины. Эти показатели получили название *центральных моментов* распределения порядка, соответствующего степени, в которую возводятся отклонения (табл. 5.8), или просто моментов (нецентральные моменты исп[illegible]дко и здесь не будут рассматриваться).

Согласно свойству средней арифметической центральный момент первого порядка равен нулю, второй центральный момент представляет собой дисперсию. Величина третьего момента μ_3 зависит, как и его знак, от преобладания положительных отклонений в кубе над отрицательными либо наоборот.

При нормальном и любом другом строго симметричном распределении сумма положительных отклонений в кубе строго равна сумме отрицательных отклонений в кубе (μ_3 используется при оценке асимметрии). Четвертый момент используется для оценки эксцесса.

Таблица 5.8

Центральные моменты

Порядок момента	Формула	
	по несгруппированным данным	по сгруппированным данным
Первый (μ_1)	$\frac{\sum_{(i)}(x_i - \bar{x})}{n}$	$\frac{\sum_{(j)}(x_j' - \bar{x})f_j}{\sum_{(j)} f_j}$
Второй (μ_2)	$\frac{\sum_{(i)}(x_i - \bar{x})^2}{n}$	$\frac{\sum_{(j)}(x_j' - \bar{x})^2 f_j}{\sum_{(j)} f_j}$
Третий (μ_3)	$\frac{\sum_{(i)}(x_i - \bar{x})^3}{n}$	$\frac{\sum_{(j)}(x_j' - \bar{x})^3 f_j}{\sum_{(j)} f_j}$
Четвертый (μ_4)	$\frac{\sum_{(i)}(x_i - \bar{x})^4}{n}$	$\frac{\sum_{(j)}(x_j' - \bar{x})^4 f_j}{\sum_{(j)} f_j}$

Показатели асимметрии

На основе момента третьего порядка можно построить показатель, характеризующий степень асимметричности распределения:

$$As = \frac{\mu_3}{\sigma^3}. \qquad (5.30)$$

Показатель *As* называют *коэффициентом асимметрии*. Он может быть рассчитан как по сгруппированным, так и по несгруппированным данным.

По данным табл. 5.6 показатель асимметрии составил:

$$As = \frac{6796}{8{,}44^3 \cdot 143} = 0{,}079,$$

т.е. асимметрия незначительна. Английский статистик К. Пирсон на основе разности между средней величиной и модой предложил другой показатель асимметрии:

$$As_П = \frac{\bar{x} - \text{Mo}}{\sigma}. \qquad (5.31)$$

По данным табл. 5.5 показатель Пирсона составил:

$$As_П = \frac{30{,}3 - 27{,}9}{8{,}44} = 0{,}284.$$

Показатель Пирсона зависит от степени асимметричности в средней части ряда распределения, а показатель асиммет-

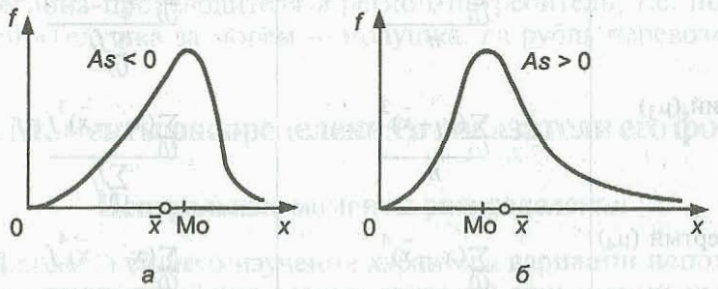


Рис. 5.3. Асимметрия распределения:
а — левосторонняя; *б* — правосторонняя

рии, основанный на моменте третьего порядка, — от крайних значений признака. Таким образом, в нашем примере в средней части распределения асимметрия более значительна, что видно и по графику (рис. 5.1). Распределения с сильной правосторонней и левосторонней (положительной и отрицательной) асимметрией показаны на рис. 5.3.

Характеристика эксцесса распределения

С помощью момента четвертого порядка характеризуется свойство рядов распределения, называемое *эксцессом*.

Показатель эксцесса рассчитывается по формуле

$$Ex = \frac{\mu_4}{4}. \qquad (5.32)$$

Часто эксцесс интерпретируется как «крутизна» распределения, но это неточно и неполно. График распределения может выглядеть сколь угодно крутым в зависимости от силы вариации признака: чем слабее вариация, тем круче кривая распределения при данном масштабе. Не говоря уже о том, что, изменяя масштабы по оси абсцисс и по оси ординат, любое распределение можно искусственно сделать «крутым» и «пологим». Для того чтобы показать, в чем состоит эксцесс деления, и правильно его интерпретировать, нужно сравнить ряды с одинаковой силой вариации (одной и той же величиной σ) и разными показателями эксцесса. Чтобы не смешать эксцесс с асимметрией, все сравниваемые ряды должны быть симметричными. Такое сравнение изображено на рис. 5.4.

Для вариационного ряда с нормальным распределением значений признака показатель эксцесса, рассчитанный по формуле (5.32), равен трем.

Однако такой показатель не следует называть термином «эксцесс», что в переводе означает «излишество». Термин «эксцесс» следует применять не к самому отношению по формуле (5.32), а к сравнению такого отношения для изучаемого распределения с величиной данного отношения для нормального распределения, т.е. с величиной 3. Отсюда окончательные формулы показателя эксцесса, т.е. излишества в сравнении с нормальным распределением при той же силе вариации, имеют вид:

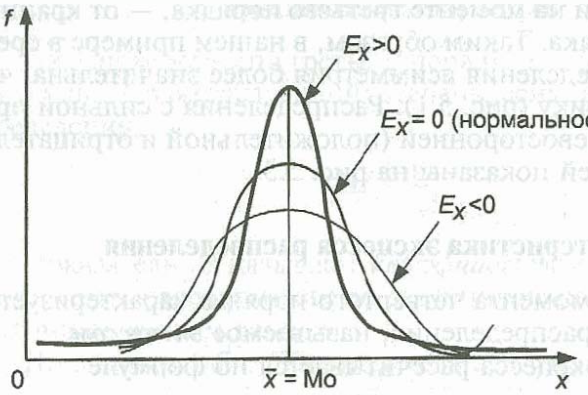


Рис. 5.4. Эксцесс распределения

для несгруппированных данных

$$Ex = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n\sigma^4} - 3; \qquad (5.33)$$

для сгруппированных данных

$$Ex = \frac{\sum_{i=1}^{n} (x_j' - \bar{x})^4 f_j}{\sigma^4 \sum_{j=1}^{n} f_j} - 3. \qquad (5.34)$$

Наличие положительного эксцесса, как и ранее отмеченного значительного различия между малым квартильным расстоянием и большим средним квадратическим отклонением, означает, что в изучаемой массе явлений существует слабо варьирующее по данному признаку «ядро», окруженное рассеянным «гало». При существенном отрицательном эксцессе такого «ядра» нет совсем.

По значениям показателей асимметрии и эксцесса распределения можно судить о близости распределения к нормальному, что бывает существенно важно для оценки результатов корреляционного и регрессионного анализа, возможностей вероятностной оценки прогнозов (гл. 9, 11). Распределение можно считать нормальным, а точнее говоря, не следует отвергать гипотезу о сходстве фактического распределения с нормальным, если показатели асимметрии и эксцесса не превышают своих двукратных средних квадратических отклонений σ_{as} и σ_{ex}. Эти средние квадратические отклонения вычисляются по формулам:

$$\sigma_{as} = \sqrt{\frac{6(n-1)\cdot n}{(n-2)\cdot(n+1)\cdot(n+3)\cdot(n+3)}}; \qquad (5.35)$$

$$\sigma_{ex} = \sqrt{\frac{24n(n-1)^2}{(n-3)\cdot(n-2)\cdot(n+3)\cdot(n+5)}}. \qquad (5.36)$$

Таким образом, если $(As/\sigma_{as}) \leqslant 2$ и $(Ex/\sigma_{ex}) \leqslant 2$, то распределение можно считать нормальным.

5.10. Предельно возможные значения показателей вариации и их применение

Применяя любой вид статистических показателей, полезно знать, каковы предельно возможные значения данного показателя для изучаемой системы и каково отношение фактически наблюдаемых значений к предельно возможным. Особенно актуальна эта проблема при изучении вариации абсолютных показателей, таких, как объем производства определенного вида продукции, наличие определенных ресурсов, распределение капиталовложений, доходов, прибыли. Рассмотрим теоретически и практически данный вопрос на примере распределения производства овощей между сельскохозяйственными предприятиями в районе.

Очевидно, что минимально возможное значение показателей вариации достигается при строго равномерном распределении объемного признака между всеми единицами совокупности, т.е. при одинаковом объеме производства в каждом из сельскохозяйственных предприятий. В таком предельном распределении (конечно, весьма маловероятном на практике) вариация отсутствует и все показатели вариации равны нулю.

Максимально возможное значение показателей вариации достигается при таком распределении объемного признака в

совокупности, при котором весь его объем сосредоточен в одной единице совокупности; например, весь объем производства овощей — в одном сельскохозяйственном предприятии района при отсутствии их производства в остальных хозяйствах. Вероятность такого предельно возможного сосредоточения объема признака в одной единице совокупности не столь уж мала; во всяком случае она гораздо больше вероятности строго равномерного распределения.

Рассмотрим показатели вариации при указанном предельном случае ее максимальности. Обозначим число единиц совокупности n, среднюю величину признака $\bar{x}$, тогда общий объем признака в совокупности выразится как $\bar{x}n$. Весь этот объем сосредоточен у одной единицы совокупности, так что $x_{max} = \bar{x}n$, $x_{min} = 0$, откуда следует, что максимальное значение амплитуды (размаха вариации) равно:

$$R_{max} = \bar{x}n - 0 = \bar{x}n, \; \rho_{max} = R_{max} : \bar{x} = n. \qquad (5.37)$$

Для вычисления максимальных значений средних отклонений абсолютного и квадратического построим таблицу отклонений (табл. 5.9).

Используя выражения, стоящие в итоговой строке табл. 5.9, получаем следующие максимально возможные значения показателей вариации.

Таблица 5.9

Модули и квадраты отклонений от средней при максимально возможной вариации

Номер единиц совокупности	Значения признака x_i	Отклонения от средней $x_i - \bar{x}$	Модули отклонений $\lvert x_i - \bar{x}\rvert$	Квадраты отклонений $(x_i - \bar{x})^2$
1	$\bar{x}n$	$\bar{x}(n-1)$	$\bar{x}(n-1)$	$\bar{x}^2(n-1)^2$
2	0	$-\bar{x}$	$\bar{x}$	$\bar{x}^2$
3	0	$-\bar{x}$	$\bar{x}$	$\bar{x}^2$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	0	$-\bar{x}$	$\bar{x}$	$\bar{x}^2$
Итого	$\bar{x}n$	0 (нуль)	$2\bar{x}(n-1)$	$\bar{x}^2[(n-1)^2 + (n-1)]$

Средний модуль отклонений, или среднее линейное отклонение:

$$a_{max} = \frac{2\bar{x}(n-1)}{n} = 2\bar{x} - \frac{2}{n}\bar{x}. \qquad (5.38)$$

Среднее квадратическое отклонение:

$$\sigma_{max} = \sqrt{\frac{\bar{x}^2[(n-1)^2 + (n-1)]}{2}} = \bar{x}\sqrt{n-1}\,. \qquad (5.39)$$

Относительное модульное (линейное) отклонение:

$$m_{max} = a_{max} : \bar{x} = 2 - \frac{2}{n}. \qquad (5.40)$$

Коэффициент вариации:

$$v_{max} = \sigma_{max} : \bar{x} = \sqrt{n-1}\,. \qquad (5.41)$$

Что касается квартильного расстояния, то система с максимально возможной вариацией обладает вырожденной структурой распределения признака, в которой не существуют («не работают») характеристики структуры: медиана, квартили и им подобные.

Полученные формулы максимально возможных значений основных показателей вариации прежде всего приводят к выводу о зависимости этих значений от объема совокупности n. Данная зависимость обобщена в табл. 5.10.

Таблица 5.10

Максимальные значения показателей вариации объемного признака при разных численностях совокупности

Численность совокупности	R	ρ	a	m	σ	v
2	$2\bar{x}$	2	$\bar{x}$	1	$\bar{x}$	1
4	$4\bar{x}$	4	$1{,}5\bar{x}$	1,5	$1{,}73\bar{x}$	1,73
6	$6\bar{x}$	6	$1{,}67\bar{x}$	1,67	$2{,}24\bar{x}$	2,24
10	$10\bar{x}$	10	$1{,}80\bar{x}$	1,80	$3\bar{x}$	3,00
20	$20\bar{x}$	20	$1{,}90\bar{x}$	1,90	$4{,}36\bar{x}$	4,36
50	$50\bar{x}$	50	$1{,}96\bar{x}$	1,96	$7\bar{x}$	7,00
100	$100\bar{x}$	100	$1{,}98\bar{x}$	1,98	$9{,}95\bar{x}$	9,95
∞	∞	∞	$2\bar{x}$	2	∞	∞

Наиболее узкие пределы изменения и слабую зависимость от численности совокупности обнаруживают средний модуль и относительное линейное отклонение. Напротив, среднее квадратическое отклонение и коэффициент вариации сильно зависят от численности единиц совокупности. Эту зависимость следует учитывать при сравнении силы интенсивности вариации в совокупностях разной численности. Если в совокупности шести предприятий коэффициент вариации объема продукции составил 0,58, а в совокупности из 20 предприятий — 0,72, то справедливо ли делать вывод о большей неравномерности объема продукции во второй совокупности? Ведь в первой, меньшей, он составил: 0,58 : 2,24 = 25,9% максимально возможного, т.е. предельного уровня концентрации производства в одном предприятии из шести, а во второй, большей, совокупности только: 0,72 : 4,36 = 16,5% максимально возможного.

Практическое значение имеет и такой показатель, как отношение фактического среднего модуля отклонений к предельно возможному. Так, для совокупности шести предприятий это соотношение составило: 0,47 : 1,67 = 0,281, или 28,1%.

Интерпретация полученного показателя такова: для перехода от наблюдаемого распределения объема продукции между предприятиями к равномерному распределению потребовалось бы перераспределить $\frac{m_{\text{факт}}}{m_{\max}} \cdot \frac{n-1}{n} = 0{,}281 \cdot \frac{6-1}{6} =$ = 0,234, или 23,4% общего объема продукции в совокупности.

Если степень фактической концентрации производства (σ или ν) составляет некоторую долю предельного значения при монополизации производства на одном предприятии, то отношение фактического показателя к предельному может характеризовать степень концентрации (или монополизации) производства.

Отношения фактических значений показателей вариации или изменения структуры к предельно возможным используются также при анализе структурных сдвигов (гл. 13).

РЕЗЮМЕ

Средние величины — важнейшие статистические показатели. При вычислении по однородным данным они характеризуют типичные значения признаков.

Показательность средней зависит не только от однородности, но и от объема данных — при прочих равных условиях чем больше объем наблюдений, тем более надежна средняя величина.

Средние, используемые статистикой, относятся к степенным средним. В зависимости от показателя степени k выделяются средние разных видов:

средняя арифметическая, $k = 1$;
средняя гармоническая, $k = -1$;
средняя квадратическая, $k = 2$;
средняя кубическая, $k = 3$;
средняя геометрическая, $k = 0$.

В соответствии со значением k величины средних образуют неравенство, называемое мажорантностью средних.

Средняя арифметическая представляет центр тяжести совокупности варьирующих значений, поскольку $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

Средняя арифметическая обладает и другими полезными свойствами.

Средняя, мода и медиана составляют показатели центра распределения. По их значениям можно сделать вывод о характере распределения — для нормального распределения: $\bar{x}$ = Me = = Mo, для распределения с правосторонней асимметрией: Mo < Me < $\bar{x}$, с левосторонней асимметрией: Mo > Me > $\bar{x}$. Для умеренно асимметричного распределения справедливо следующее соотношение: $\bar{x}$ ближе к Me, нежели к Mo.

Средние подразделяются на простые и взвешенные. Взвешивание позволяет отразить реальное значение отдельных вариант. Чем сильнее варьируют веса и чем сильнее корреляция между осредняемым признаком и весом, тем больше значение взвешенной средней отличается от значения простой средней, рассчитанной по тем же данным.

При большом числе наблюдений среднее значение и показатели вариации рассчитываются по вариационному ряду. Вид вариационного ряда зависит от вида варьирующего признака: дискретный или непрерывный.

Большое значение в анализе данных имеют кумулятивные распределения: «больше, чем» и «не меньше, чем».

При группировке с неравными интервалами взвешивание проводится по плотности распределения.

Медиана и мода относятся к структурным характеристикам ряда распределения, так же как и децили, квартили, квинтили.

Размер и интенсивность вариации измеряются следующими показателями: размах вариации, среднее линейное отклонение от средней (среднее абсолютное отклонение), среднее квадратическое отклонение, дисперсия, коэффициент вариации. Если значение среднего квадратического отклонения составляет половину и более значения средней, то данные можно считать неоднородными.

Для оценки точности расчетов по вариационному ряду можно применить правило сложения дисперсий. Общая дисперсия равна сумме межгрупповой и внутригрупповой дисперсий. Чем меньше величина внутригрупповой дисперсии, чем ближе середины интервалов переменной x к величинам групповых средних, тем точнее расчеты по вариационному ряду, тем они ближе к результатам расчетов по несгруппированным данным. Особенно это следует принимать во внимание при расчете дисперсии.

Показатели асимметрии распределения и эксцесса дают представление о характере распределения: $As > 0$ — правосторонняя асимметрия, $As < 0$ — левосторонняя асимметрия. Для нормального распределения $As = 0$. Положительное значение эксцесса ($Ex > 0$) свидетельствует о крутизне распределения (однородности), отрицательное ($Ex < 3$) — о пологости, разнородности данных. Для нормального распределения $Ex = 3$.

Имеет смысл сравнивать показатели вариации не только с характеристиками нормального распределения, но и с предельно возможными значениями при данной численности наблюдений.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Джини К.* Средние величины. — М.: Статистика, 1970.
2. *Кривенкова Л. Н., Юзбашев М. М.* Область существования показателей вариации и ее применение // Вестник статистики. — 1991. — № 6. — С. 66—70.
3. *Макарова Н. В., Трофимец В. Я.* Статистика в Excel. — М.: Финансы и статистика, 2002.
4. *Пасхавер И. С.* Средние величины в статистике. — М.: Статистика, 1979.
5. *Тюрин Ю. Н., Макаров А. А.* Анализ данных на компьютере. — М.: Финансы и статистика. — Инфра-М, 1995.

Глава 6

ГРУППИРОВКА

6.1. Значение и сущность группировки

Русский статистик Д. П. Журавский (1810—1856) очень точно определил статистику как «счет по категориям». Действительно, среди бесконечного разнообразия явлений мы, как правило, улавливаем наличие некоторого конечного числа групп или типов.

Лицо каждого человека неповторимо, и все-таки можно классифицировать лица по типам (скуластое, продолговатое, круглое); предприятия образуют группы по формам собственности, характеру производимой продукции, размерам (крупные, средние, мелкие), финансовому положению; государства делятся на группы по уровню экономического развития и т.д. Примеров можно приводить много, но ясно, что какую бы совокупность мы ни изучали, она всегда подразделяется на группы. Это обусловлено такими объективными свойствами явлений, как вариация, наличие частных совокупностей (см. гл. 1).

Группировка — это распределение единиц по группам в соответствии со следующим принципом: различия между единицами, отнесенными к одной группе, должны быть меньше, чем между единицами, отнесенными к разным группам.

Группировка лежит в основе дальнейшей работы с собранной информацией. На основе группировки рассчитываются сводные показатели по группам, появляется возможность их сравнения, анализа причин различий между группами, изучения взаимосвязей между признаками. Если рассчитать сводные показатели только в целом по совокупности, то мы не сможем уловить ее структуры, роли отдельных групп, их специфики.

Например, можно рассчитать среднюю прибыль на одно предприятие, обобщая данные по всем предприятиям данной территории, а можно первоначально разделить их на прибыльные и убыточные, прибыльные — на подгруппы по величине прибыли и только после этого приступить к расчетам средней прибыли в каждой группе (для убыточных предприятий финансовый результат — это средняя сумма убытка на одно предприятие). Тогда можно сравнить успешность работы предприятий по группам, узнать долю каждой группы в общей численности предприятий. Очевидно, что дифференцированный подход даст больше информации и обеспечит лучшее качество анализа и выводов.

Однородность (гомогенность) данных является исходным условием их статистического описания и анализа — вычисления и интерпретации обобщающих показателей, построения уравнения регрессии, измерения корреляции (гл. 9), статистического умозаключения (гл. 7, 8).

Таким образом, значение группировки состоит в том, что этот метод обеспечивает обобщение данных, представление их в компактном, обозримом виде. Кроме того, группировка создает основу для последующей сводки и анализа данных.

Для изучения структурных изменений в экономике государственная статистика использует группировку хозяйственных субъектов по формам собственности и организационно-правовым формам (табл. 6.1).

Сводные показатели для отдельных групп являются типичными и устойчивыми, если, во-первых, группировка проведена правильно, во-вторых, группы имеют достаточную численность. Первое условие связано с тем, что деление на группы далеко не всегда очевидно. Выполнение второго условия необходимо, так как при достаточно большом числе единиц (не менее пяти в группе) в сводных показателях взаимопогашаются случайные характеристики и проявляются закономерные, типичные.

Для решения задачи группировки нужно установить правила отнесения каждой единицы к той или иной группе.

В эти правила входят определения тех характеристик (признаков), по которым будет проводиться группировка (так называемых *группировочных признаков*), и их значений, отделяющих одну группу от другой (*интервалы группировки*).

Таблица 6.1

Организационно-правовые формы и формы собственности хозяйственных субъектов Российской Федерации

Организационно-правовые формы	Форма собственности						
	государственная	муниципальная	общественных объединений	частная	смешанная (без иностранного участия)	собственность иностранных юридических лиц, граждан и лиц без гражданства	смешанная с российским и иностранным капиталом
Государственные предприятия							
Муниципальные предприятия							
Акционерные общества и товарищества							
Из них:							
акционерные общества открытого типа							
акционерные общества закрытого типа							
товарищества с ограниченной ответственностью							
смешанные товарищества							
полные товарищества							
Сельскохозяйственные предприятия (организации)							
Хозяйства населения							
Крестьянские (фермерские) хозяйства							
Объединения (союзы, ассоциации) крестьянских (фермерских) хозяйств							
Производственные кооперативы							
Предприятия потребительской кооперации							
Предприятия общественных и религиозных организаций							
Индивидуальные (семейные) частные предприятия (с привлечением наемного труда)							

Продолжение

Организационно-правовые формы	Форма собственности						
	государственная	муниципальная	общественных объединений	частная	смешанная (без иностранного участия)	собственность иностранных юридических лиц, граждан и лиц без гражданства	смешанная с российским и иностранным капиталом
Другие предприятия							
Другие объединения предприятий							
Некоммерческие организации:							
государственные учреждения							
муниципальные учреждения							
общественные и религиозные организации, их учреждения							
потребительские общества, учреждения потребительской кооперации							
фонды							
другие некоммерческие организации							
Филиалы, представительства и другие обособленные подразделения							

Группировка называется *простой* (монотетической), если для ее построения используется один группировочный признак. Если группировка проводится по нескольким признакам, она называется *сложной* (политетической). Обычно такая группировка проводится как *комбинационная*, т.е. группы, выделенные по одному признаку, подразделяются на подгруппы по другому признаку. Казалось бы, этот метод выделения групп должен быть лучше простой группировки — ведь трудно ожидать, что различия между группами можно уловить лишь на основе одного признака. Однако комбинация признаков приводит к дроблению совокупности в геометрической прогрессии: число групп будет равно произведению числа вы-

деленных категорий по каждому из них. Так, если по одному признаку выделено l категорий, а по второму — m, то общее число категорий составит: $k = l \cdot m$. Данные становятся труднообозримыми, группы включают малое число единиц, групповые показатели делаются ненадежными.

Альтернативой является проведение *многомерных группировок*, или многомерных классификаций (подразд. 6.3).

Остановимся на определении интервалов группировочных признаков. Используются интервалы *открытые* и *закрытые*. В последнем случае указываются *верхняя* и *нижняя границы интервала*. Например, группы крупных и средних предприятий по численности работников, человек: 200—600, 600—1000, 1000—2000. Такая запись предполагает, что единица, у которой значение признака совпадает с верхней границей интервала, относится к следующей группе, т.е. интервал читается как «от — до».

Иногда границы закрытых интервалов предполагают включение единиц с нижней и верхней границами. Например, группировка населения по возрасту, лет: 0—4, 5—9, 10—14, 15—19, 20—24, 25—29 и т.д. Интервал называется открытым, если указана либо только верхняя, либо только нижняя граница: до 200 человек или 2000 человек и более.

Закрытые интервалы подразделяются на *равные* и *неравные*. Как указывалось в гл. 5, величина равного интервала находится по формуле

$$i_x = \frac{x_{\max} - x_{\min}}{\text{число групп}}.$$

Неравные интервалы могут определяться как *равнонаполненные*. При этом совокупность разделяется на группы равного объема с числом единиц в каждой j-й группе: $n_j = n : m$, где n — общее число единиц; m — число групп. Данные ранжируются, отсчитывается число единиц, составляющих первую группу n_1, затем — вторую n_2 и т.д. Границы интервалов будут соответствовать фактическим значениям признака в каждой группе.

Бывает, что число групп заранее неизвестно и определяется опытным путем на основе перебора вариантов группировки, выявления такого варианта, который наилучшим образом позволяет увидеть различия между группами.

При определении числа групп следует обращать внимание на то, чтобы в одну группу не попало свыше половины всех единиц совокупности.

Если группировочный признак неколичественный, или количественный дискретный с малым числом значений, то группировка данных проводится путем подсчета числа единиц с данным значением признака. Примером такой группировки является табл. 6.2.

Таблица 6.2

Группировка станкостроительных заводов по числу производимых типов станков

Число типов станков	1	2	3	4	5 и более
Число заводов	19	10	7	3	1

Очевидно, что метод группировок тесно связан с представлением данных в виде групповых, или комбинационных, таблиц, а также с графиками структуры совокупности, ее частей и соотношений между ними.

6.2. Виды группировок

Группировка проводится с целью установления статистических связей и закономерностей, построения описания объекта, выявления структуры изучаемой совокупности. Различия в целевом назначении группировки выражаются в существующей в отечественной статистике классификации группировок: типологические, структурные, аналитические.

Типологическая группировка служит для выделения социально-экономических типов. Этот вид группировок в значительной степени определяется представлениями экспертов о том, какие типы могут встретиться в изучаемой совокупности.

Чтобы пояснить особенность этой группировки, приведем последовательность действий для ее проведения:

1) называются те типы явлений, которые могут быть выделены;

2) выбираются группировочные признаки, формирующие описание типов;

3) устанавливаются границы интервалов;

4) группировка оформляется в таблицу, выделенные группы (на основе комбинации группировочных признаков) объединяются в намеченные типы, и определяется численность каждого из них.

Пример. Поставлена задача выделить типы акционерных компаний с высокими, средними и низкими дивидендами и установить распространенность каждого типа в данном регионе.

Показатель выплаты дивидендов характеризует долю прибыли на акцию или долю чистого дохода, выплачиваемого как дивиденды.

$$\text{Показатель выплаты дивидендов} = \frac{\text{Дивиденды}}{\text{Чистый доход}} \cdot 100\%.$$

Этот коэффициент зависит от структуры акционерного капитала фирмы, длительности существования фирмы и перспектив ее роста. Обычно молодые, быстрорастущие компании выплачивают низкие дивиденды, если вообще их выплачивают; тогда как компании, давно работающие на рынке, стремятся дать более высокие дивиденды. Структура капитала и выплата дивидендов зависят от отраслевой принадлежности фирмы. Поэтому при классификации фирм по уровню выплаты дивидендов мы должны использовать в качестве группировочных признаков, во-первых, отрасль (подотрасль), во-вторых, показатель выплаты дивидендов.

Первый группировочный признак выполняет роль характеристики условий, второй непосредственно характеризует тип фирмы. Границы интервалов для второго группировочного признака могут изменяться при переходе от одной отрасли к другой, так как то, что для одной отрасли может рассматриваться как высокий уровень выплаты, для другой может оцениваться иначе.

Изменение границ интервалов группировочного признака при выделении одних и тех же типов в разных условиях называется *специализацией интервалов группировочного признака.*

Иногда условия формирования типов приводят к различиям в их описании, в самих признаках. Например, выделение крупных, средних, мелких предприятий в разных отраслях должно проводиться по разным характеристикам: в энергоемких отраслях — по потреблению электроэнергии; в сырьеемких — по величине товарно-материальных запасов; в трудоемких — по численности рабочих; в капиталоемких — по стоимости оборудования.

Изменение круга группировочных признаков при выделении одних и тех же типов в разных условиях называется *специализацией группировочных признаков.*

Вернемся к нашему примеру. Предположим, что мы располагаем данными 15 фирм, представляющих три подотрасли промышленности. Проведем их группировку с учетом двух выше названных признаков (табл. 6.3).

Таблица 6.3

Группировка акционерных компаний *n*-го района по уровню выплаты дивидендов за 200_ г.

Подотрасль промышленности	Показатель выплаты дивидендов, %	Тип компании	Число компаний
Производство детских игрушек	до 30	н	—
	30—50	с	1
	50 и выше	в	4
Производство животного масла	до 20	н	1
	20—40	с	2
	40 и выше	в	—
Производство хлопчатобумажных тканей	до 10	н	2
	10—30	с	4
	30 и выше	в	1

Примечание. Здесь: н — с низким показателем выплаты дивидендов; с — со средним показателем выплаты дивидендов; в — с высоким показателем выплаты дивидендов.

Использование специализации интервалов как бы уравнивает наши оценки компаний в разных отраслях, что позволяет объединить выделенные группы в три типа независимо от отрасли (табл. 6.4). Это последний шаг типологической группировки.

Как видим, данный метод позволяет избежать чрезмерного дробления совокупности, но он слишком субъективен: эксперт определяет, какие типы должны быть выделены, по каким признакам, какими должны быть границы интервалов. К тому же число группировочных признаков ограничено дву-

Таблица 6.4

Распределение акционерных компаний *n*-го района по типам в 200_ г.

Тип компании	Число компаний	
	абсолютное	% к итогу
н	3	20,0
с	7	46,7
в	5	33,3
Итого	15	100,0

мя-тремя. Однако если объект исследования хорошо изучен, если имеется развитая теория, то этот метод может дать хорошо интерпретируемые результаты.

В любом случае правильность проведения типологической группировки требует проверки. С этой целью рассчитываются сводные показатели по группам (средние, относительные величины); если различие между группами *статистически незначимо* (по *t*-критерию Стьюдента или *F*-критерию, или критерию χ^2 и т.д. (гл. 7)), то схема группировки должна быть пересмотрена — схожие группы могут быть объединены, изменены границы интервалов и т.д.

Структурная группировка характеризует структуру совокупности по какому-либо одному признаку (табл. 6.5).

Если для типологической группировки чаще используются открытые и неравные интервалы, то для структурной группировки более характерны закрытые равные интервалы. Структурная группировка — это ряд распределения. Она позволяет изучать интенсивность вариации группировочного признака (см. гл. 5). На основе структурной группировки можно изучать динамику структуры совокупности.

Если известны структурные характеристики совокупности в одном и другом периодах: w_{i0} и w_{i1} — доли i-й группы в период «0» и в период «1», то можно рассчитать показатель среднего абсолютного изменения структуры:

$$d_{w_1-w_0} = \frac{\sum_{i=1}^{k} |w_{1i} - w_{0i}|}{k}, \qquad (6.1)$$

где k — число групп.

Таблица 6.5

Распределение крестьянских (фермерских) хозяйств России по размеру земельного участка (на конец года; в процентах)

Размер земельного участка, га	1995	2001
0	0,4	1,1
3 и менее	12,2	16,9
4—5	10,4	9,7
6—10	15,0	13,9
10—20	18,5	15,5
21—50	22,3	18,7
51—70	6,8	6,0
71—100	6,1	5,7
101—200	5,8	7,0
свыше 200	2,5	5,5
Итого	100	100

Источник. Россия в цифрах, 2002. Краткий статистический сборник. — М.: Госкомстат России, 2002. — С. 144—165.

По данным табл. 6.5 структура крестьянских (фермерских) хозяйств по размеру земельного участка изменилась в среднем на 1,92 процентного пункта на группу.

Другой сводный показатель абсолютных структурных сдвигов строится на основе формулы среднего квадратического отклонения:

$$s_{w_1-w_0} = \sqrt{\frac{\sum_{i=1}^{k} (w_{1i} - w_{0i})^2}{k}}. \qquad (6.2)$$

Если показатели структуры выразить не в долях, а в процентах, то, так же как и первый показатель, квадратичный коэффициент абсолютных структурных сдвигов оценивает: на сколько процентных пунктов в среднем различаются удельные веса групп сравниваемых структур. При отсутствии структурных сдвигов оба эти показателя равны нулю; их величина тем больше, чем значительнее абсолютные изменения удельных весов групп. Квадратичный коэффициент более чутко реагирует на структурные изменения. Существуют и

другие показатели для измерения структурных сдвигов (см., например, индекс структуры в гл. 14). При сравнениях предполагается, что число групп в одном и другом периодах остается одним и тем же. По данным табл. 6.5 $s_{w_1 - w_0} = 2{,}4$ процентного пункта, т.е. средний квадратичный показатель превышает средний арифметический (по свойству мажорантности средних).

Деление группировок на типологические и структурные достаточно условно. Если задать, например, границы среднедушевого дохода, соответствующие определенным типам благосостояния, то можно с полным правом назвать полученную группировку типологической.

Аналитическая группировка характеризует взаимосвязь между двумя и более признаками, из которых один рассматривается как результат, другой (другие) — как фактор (факторы).

Пример однофакторной аналитической группировки представлен в табл. 6.6.

В данном примере оборачиваемость в днях — фактор, обозначенный x, прибыль — результат — y. Очевидно, что при одной и той же продолжительности оборота предприятия могут иметь разную прибыль. Для того чтобы установить связь между признаками, данные группируются по признаку-фактору. Затем по каждой группе рассчитывается среднее значение результата. По обобщенным данным гораздо легче увидеть, есть связь между признаками или нет, прямая связь или обратная, линейная или нелинейная. Эти выводы делаются

Таблица 6.6

Характеристика зависимости прибыли предприятий от оборачиваемости оборотных средств за 200_ г.

Продолжительность оборота средств, дней	Число предприятий	Середина интервала, дней	Средняя прибыль, млн руб.	Изменение средней прибыли, млн руб.
x_j	n_j	x_j'	$\bar{y}_j$	$\bar{y}_j - \bar{y}_{j-1}$
11—30	6	20	14,60	—
31—50	8	40	12,95	−1,65
51—70	6	60	7,40	−5,55
Итого	20	x	11,78	x

путем сопоставления изменений средних значений результата по группам с изменениями фактора ($\bar{y}_j - \bar{y}_{j-1}$). Для того чтобы эти изменения были сравнимыми, следует проводить группировку с равными интервалами или рассчитывать изменения результата на единицу изменений фактора.

В нашем примере средняя прибыль изменяется от группы к группе, следовательно, связь между оборачиваемостью и прибылью существует, причем обратная: чем медленнее оборачиваются оборотные средства, тем меньше прибыль.

Рассчитаем, насколько снижается прибыль при замедлении оборачиваемости от 11—30 до 31—50 дней и при замедлении оборачиваемости от 31—50 до 51—70 дней:

$$1)\ b_{yx} = \frac{\bar{y}_2 - \bar{y}_1}{x_2' - x_1'} = \frac{\Delta \bar{y}}{\Delta x'} = \frac{-1{,}65}{20} = -0{,}0825 \text{ млн руб./день,}$$

или −82,5 тыс. руб./день;

$$2)\ b_{yx} = \frac{\bar{y}_3 - \bar{y}_2}{x_3' - x_2'} = \frac{\Delta \bar{y}}{\Delta x'} = \frac{-5{,}55}{20} = -0{,}2775 \text{ млн руб./день,}$$

или −277,5 тыс. руб./день.

Полученные значения показывают среднюю величину снижения прибыли при замедлении оборачиваемости на один день. Такие показатели называются показателями *силы связи*. Различие в их значениях свидетельствует, что сила влияния оборачиваемости на прибыль не является постоянной — она возрастает при сроках оборачиваемости свыше 50 дней, т.е. можно предположить, что связь признаков нелинейная.

В случае линейной связи важным показателем является характеристика *средней силы связи*:

$$b_{yx} = \frac{\bar{y}_m - \bar{y}_1}{x_m' - x_1'}, \qquad (6.3)$$

где $\bar{y}_m$, $\bar{y}_1$ — средние значения результативного признака в последней и первой группах соответственно;

x_m', x_1' — середины интервалов (или средние значения) факторного признака в последней и первой группах.

В случае прямой связи — $b_{yx} > 0$, в случае обратной связи — $b_{yx} < 0$.

По данным табл. 6.6

$b_{yx} = \frac{7,4 - 14,67}{60 - 20} = -0,180$ млн руб./день, или −180 тыс. руб./день.

При нелинейной связи показатель средней силы связи не рассчитывается.

По аналитической группировке можно измерить связь с помощью еще одного показателя: *эмпирического корреляционного отношения*. Этот показатель обозначается греческой буквой η (эта). Он основан на *правиле разложения дисперсии*, согласно которому *общая дисперсия s_y^2 равна сумме внутригрупповой и межгрупповой дисперсий.*

Дисперсия результативного признака внутри группы при относительном постоянстве признака-фактора возникает за счет других факторов (не связанных с изучаемым). Такая дисперсия называется *остаточной* (та колеблемость, которая осталась при закреплении изучаемого фактора x). Это внутригрупповая дисперсия. Она определяется по формуле

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j}, \qquad (6.4)$$

где y_{ij} — значение признака y для i-й единицы в j-й группе;

$j = 1, 2, 3, ..., m$;

$\bar{y}_j$ — среднее значение признака y в j-й группе;

n_j — число единиц в j-й группе.

Внутригрупповые дисперсии, рассчитанные для отдельных групп, объединяются в средней величине внутригрупповой дисперсии:

$$\overline{\sigma_{yx}^2} = \frac{\sum_{1}^{m} \sigma_j^2 n_j}{\sum_{1}^{m} n_j}. \qquad (6.5)$$

Межгрупповая дисперсия относится на счет изучаемого фактора (и факторов, связанных с ним), поэтому называется *факторной*. Она определяется по формуле

$$\sigma_{\bar{y}_x}^2 = \frac{\sum_{1}^{m} (\bar{y}_j - \bar{y})^2 n_j}{\sum_{1}^{m} n_j}. \qquad (6.6)$$

Правило сложения дисперсий (или разложения дисперсии) может быть записано:

$$\sigma_y^2 = \overline{\sigma_{yx}^2} + \sigma_{\bar{y}_x}^2 \qquad (6.7)$$

или

$$\sum_{(j)}\sum_{(i)} (y_{ij} - \bar{y})^2 = \sum_{(j)}\sum_{(i)} (y_{ij} - \bar{y}_j)^2 + \sum_{(j)} (\bar{y}_j - \bar{y})^2 n_j. \qquad (6.8)$$

Эмпирическое корреляционное отношение измеряет, какую часть общей колеблемости результативного признака вызывает изучаемый фактор. Соответственно этот показатель рассчитывается на основе отношения факторной дисперсии к общей дисперсии результативного признака:

$$\eta^2 = \frac{\sigma_{\bar{y}_x}^2}{\sigma_y^2} \text{ — коэффициент детерминации;} \qquad (6.9)$$

$$\eta = \sqrt{\frac{\sigma_{\bar{y}_x}^2}{\sigma_y^2}} \text{ — эмпирическое корреляционное отношение.} \qquad (6.10)$$

Этот показатель принимает значения в интервале [0, 1]: чем ближе к 1, тем теснее связь, и наоборот. Можно руководствоваться рекомендациями по оценке тесноты связи: если $\eta \leqslant 0,3$ — связь слабая; $0,3 < \eta \leqslant 0,5$ — связь заметная; $0,5 < \eta \leqslant 0,7$ — связь умеренно тесная; $\eta > 0,7$ — связь тесная.

В нашем примере: $\sum_{(i)} (y_i - \bar{y})^2 = 440,6$; $\sum_{(j)}\sum_{(k)}\sum_{(i)} (y_{ijk} - \bar{y}_{jk})^2 = 266,83$; $\sum_{(j)}\sum_{(k)} (\bar{y}_{i_k} - \bar{y})^2 = 173,77$.

Отсюда корреляционное отношение равно: $\eta = \sqrt{\frac{173{,}77}{440{,}6}} = 0{,}628$ — связь умеренно тесная.

Для изучения влияния нескольких факторов на результат строится *многофакторная аналитическая группировка* как комбинационная группировка по признакам-факторам, и для каждой подгруппы рассчитывается среднее значение результативного признака.

Обратимся к рассмотренному выше примеру, который дополним вторым фактором формирования прибыли — величиной запаса оборотных средств (z); по этому фактору выделены три группы (табл. 6.7).

Эта группировка позволяет проследить колеблемость прибыли под влиянием двух факторов. Конечно, уверенность нашего заключения в том, что прибыль изменяется от группы к группе именно за счет изменений запаса оборотных средств и скорости их обращения, зависит от того, насколько обеспечено погашение влияния прочих факторов, т.е. от числа единиц в подгруппах (n_{jk}). В данном примере наполненность групп недостаточна для того, чтобы выявить «чистое» влияние изучаемых факторов.

Таблица 6.7

Характеристика зависимости прибыли предприятий от величины запаса и оборачиваемости оборотных средств за 200_ г.

Средний запас оборотных средств, млн руб.	Оборачиваемость, дней	Число предприятий	Средняя прибыль, млн руб.	Колеблемость прибыли по группам
z_k	x_j	n_{jk}	$\bar{y}_{jk}$	$(\bar{y}_{jk} - \bar{y})^2 n_{jk}$
55—85	11—30	1	14,300	6,3504
	31—50	2	10,650	2,5538
	51—70	1	7,300	20,0104
85—115	11—30	2	12,150	0,2738
	31—50	4	12,075	0,3481
	51—70	2	6,850	48,6098
115—145	11—30	3	16,333	62,1894
	31—50	2	17,000	54,4968
	51—70	3	7,800	47,5212
Итого	x	20	11,780	242,3537

При $n_{jk} \geqslant 5$ многофакторная аналитическая группировка позволяет измерить силу связи между результатом и одним из факторов при постоянстве второго фактора, т.е. получить так называемые *частные* (или *чистые*) показатели силы связи.

По данным табл. 6.7 рассчитаны показатели силы связи между прибылью и оборачиваемостью при закреплении уровня запасов оборотных средств. Таких показателей три (по числу групп по фактору z):

$$b_{yx,z_1} = \frac{7{,}30 - 14{,}30}{60 - 20} = -0{,}175 \text{ млн руб./день},$$

$$b_{yx,z_2} = \frac{6{,}85 - 12{,}15}{60 - 20} = -0{,}132 \text{ млн руб./день},$$

$$b_{yx,z_3} = \frac{7{,}80 - 16{,}333}{60 - 20} = -0{,}213 \text{ млн руб./день}.$$

Точно так же могут быть вычислены показатели силы связи между прибылью и запасом оборотных средств при закреплении оборачиваемости:

$$b_{yz,x_1} = \frac{16{,}333 - 14{,}3}{130 - 70} = 0{,}034 \text{ млн руб./млн руб.},$$

$$b_{yz,x_2} = \frac{17{,}0 - 10{,}65}{130 - 70} = 0{,}106 \text{ млн руб./млн руб.},$$

$$b_{yz,x_3} = \frac{7{,}8 - 7{,}3}{130 - 70} = 0{,}008 \text{ млн руб./млн руб.}$$

В данном примере чистое влияние первого фактора (обратное) возрастает при увеличении уровня закрепленного фактора (величины запасов оборотных средств), а чистое влияние второго фактора снижается.

Можно рассчитать и показатель множественной тесноты связи — *совокупное эмпирическое корреляционное отношение*. Для трех признаков, как в нашем примере, его формула следующая:

$$\eta_{yxz} = \sqrt{\frac{\delta^2_{y_{xz}}}{\sigma^2_y}}, \quad (6.11)$$

Так же как и показатель парной связи, η_{yxz} принимает значение в интервале [0; 1]. В числителе подкоренного выражения находится факторная дисперсия результативного признака:

$$\delta^2_{\bar{y}_{xz}} = \frac{\sum\sum_{(k)(j)} (\bar{y}_{jk} - \bar{\bar{y}})^2 n_{jk}}{\sum\sum_{(k)(j)} n_{jk}}. \qquad (6.12)$$

По данным табл. 6.7

$$\eta_{yxz} = \sqrt{242{,}3537/440{,}6} = \sqrt{0{,}550} = 0{,}742.$$

Можно с некоторыми оговорками заключить, что на 55% вариация прибыли в этой совокупности предприятий определяется вариацией изучаемых факторов.

Многофакторная аналитическая группировка — очень гибкий прием изучения связей. Она позволяет уловить влияние факторов на результат с изменением условий (закреплением прочих факторов на разных уровнях).

Однако при всех отмеченных плюсах этот метод имеет огромный минус — дробление совокупности, в результате чего выделяются подгруппы с малым числом единиц. В этом случае средние значения результативного признака неустойчивы, не достигается погашение прочих факторов, соответственно ненадежными становятся и показатели связи. Но если совокупность большого объема и распределение признаков-факторов не являются крайне асимметричными, этот метод, как никакой другой, позволяет получить много информации об отношениях между переменными.

В какой-то мере избежать дробление данных и при этом получить «чистые» характеристики связей между переменными позволяет применение *метода стандартизации распределений* в комбинационной таблице. Если в группах по одной переменной, скажем по z в табл. 6.7, распределение по другой переменной x принять стандартным и на его основе рассчитать групповые средние величины результативного признака, то они будут отличаться за счет принадлежности к разным группам по признаку z при элиминировании признака x. В качестве стандартного применяется распределение в целом по совокупности. Так, по данным табл. 6.7 стандартное распределение по x следующее:

x_1 — 6 ед., x_2 — 8 ед., x_3 — 6 ед., или в относительном выражении — 0,3; 0,4; 0,3. Тогда средняя прибыль при заданном значении переменной z при стандартизации распределения по переменной x (стандартизованная средняя) равна:

в первой группе

$$\bar{y}_{1\text{станд}} = \sum_j \bar{y}_{j1} w_j = 14{,}3 \cdot 0{,}3 + 10{,}65 \cdot 0{,}4 + 7{,}3 \cdot 0{,}3 = 10{,}74;$$

во второй группе

$$\bar{y}_{2\text{станд}} = \sum_j \bar{y}_{j2} w_j = 12{,}15 \cdot 0{,}3 + 12{,}075 \cdot 0{,}4 + 6{,}85 \cdot 0{,}3 = 10{,}53;$$

в третьей группе

$$\bar{y}_{3\text{станд}} = \sum_j \bar{y}_{j3} w_j = 16{,}333 \cdot 0{,}3 + 17{,}0 \cdot 0{,}4 + 7{,}8 \cdot 0{,}3 = 14{,}04.$$

На основе полученных стандартизованных средних рассчитаем показатель «чистой» связи между величиной прибыли и средним запасом оборотных средств.

$$b_{yz(\text{станд})} = \frac{14{,}04 - 10{,}74}{130 - 70} = 0{,}055 \text{ млн руб./млн руб.}$$

Попробуйте сделать такой же расчет для второго фактора. Стандартизация распределения по переменной z, расчет стандартизованных средних результативного признака и показателей «чистой» связи между y и x при элиминировании z проводятся аналогично. Заметим, что рассмотренные приемы анализа не входят пока в ППП для ПЭВМ.

В современных средствах программного обеспечения группировкам не уделяется должного внимания. А ведь группировки — действительно гибкий инструмент анализа. Но, отдавая должное этому методу, следует помнить, что надежность выводов на их основе зависит от объема совокупности (n) и количества выделенных групп (m). Отмечая возможность получения субъективных выводов на основе группировок, русский статистик А. А. Чупров (1873—1926) писал: «...нередко случается, что ловкой обработкой одного и того же материала можно выжать из него при помощи этого приема прямо противоположные заключения».

Таблица 6.8

Зависимость среднемесячной добычи угля от длины лавы

Длина лавы, м	Число забоев	Среднемесячная добыча угля, т	Длина лавы, м	Число забоев	Среднемесячная добыча угля, т	Длина лавы, м	Число забоев	Среднемесячная добыча угля, т
x	n_j	$\bar{y}_j$	x	n_j	$\bar{y}_j$	x	n_j	$\bar{y}_j$
22—50	4	1435	22—57	11	1720	23—69	16	2416
50—78	14	2567	57—92	36	3015	69—115	60	3274
78—106	54	3177	92—127	60	2986	115—161	42	2991
106—134	35	2644	127—162	11	4948			
134—162	11	4968						
Итого	118	3060	Итого	118	3060	Итого	118	3060

Чтобы избежать этого, необходимо руководствоваться следующими правилами. Прежде всего выделенные группы не должны содержать по одной единице. В группе должно быть не менее пяти единиц. Нельзя выделять лишь две группы, так как между двумя группами обязательно будут различия, которые трудно объяснить. Так что минимальное количество групп равно $m_{\min} = 3$.

Рассмотрим влияние количества групп на выводы о характере связи между признаками.

Пример. С целью изучения зависимости среднемесячной добычи угля (y) от длины лавы (x) по совокупности 118 забоев, оборудованных угольными комбайнами, проведены три варианта группировки с разным числом групп: 5, 4 и 3. Результаты группировок представлены в табл. 6.8.

Возникает вопрос: какой же из вариантов группировки лучше раскрывает связь между длиной лавы в забое и добычей угля? По-видимому, можно попытаться ответить с помощью корреляционного отношения или коэффициента детерминации: где доля межгрупповой дисперсии выше, та группировка и будет лучшей. Значения корреляционного отношения оказались следующими:

Число групп	Корреляционное отношение
$m = 5$	$\eta = 0{,}409$
$m = 4$	$\eta = 0{,}395$
$m = 3$	$\eta = 0{,}159$

Поскольку при трех группах величина корреляционного отношения резко снизилась, то можно отклонить этот вариант группировки, так как он не раскрывает влияние длины лавы на добычу угля. Поэтому остается выбрать между первой и второй группировками. По величине корреляционного отношения можно заключить, что группировка на пять групп лучше. Но окончательный вывод можно сделать на основе соотношения величины корреляционного отношения (η_{yx}) с его *случайным значением* (η_0), которое зависит от числа групп и числа наблюдений:

$$\eta_0 = \sqrt{\frac{m-1}{n-1}}.$$

Группировка, для которой отношение ($\eta_{yx} : \eta_0$) больше, признается лучшей. В данном случае при пяти группах $\eta_{yx} : \eta_0 = 2{,}21$; при четырех группах $\eta_{yx} : \eta_0 = 2{,}47$.

В рассмотренном примере группировка с четырьмя группами максимально раскрывает действие признака-фактора на результат.

6.3. Многомерные группировки

Мы убедились, как трудно выбрать какой-то один признак в качестве основы группировки. Еще труднее проводить группировку по нескольким признакам. Комбинация двух признаков позволяет сохранить обозримость таблицы, но комбинация трех или четырех признаков дает совершенно неудовлетворительный результат: ведь даже при выделении трех категорий по каждому из группировочных признаков мы получим 9 или 27 подгрупп. Равномерность распределения единиц по группам в принципе невозможна. Вот и получаются группы, в которые входят 1—2 наблюдения. Сохранить сложность описания групп и вместе с тем преодолеть недостатки комбинационной группировки позволяют *методы многомерных группировок*. Часто их называют *методами многомерной классификации*.

Эти методы получили распространение благодаря использованию ПЭВМ и пакетов прикладных программ. Цель этих методов — классификация данных, иначе говоря, группиров-

ка на основе множества признаков. Такие задачи широко распространены в науках о природе и обществе, в практической деятельности по управлению массовыми процессами. Например, выделение типов предприятий по финансовому положению, по экономической эффективности деятельности проводится на основе множества признаков; то же при выделении групп клиентов в банке.

Простейшим вариантом многомерной классификации является группировка на основе многомерных средних.

Многомерной средней называется средняя величина нескольких признаков для одной единицы совокупности. Поскольку нельзя рассчитать среднюю величину абсолютных значений разных признаков, выраженных в разных единицах измерения, то многомерная средняя вычисляется из относительных величин, как правило, — из отношений значений признаков для единицы совокупности к средним значениям этих признаков:

$$\bar{p}_i = \frac{\sum_{j=1}^{k} p_{ij}}{k} = \sum_{j=1}^{k} \left(\frac{x_{ij}}{\bar{x}_j}\right) : k, \qquad (6.13)$$

где $\bar{p}_i$ — многомерная средняя для i-й единицы;
i — номер единицы совокупности;
j — номер признака;
k — число признаков;
x_{ij} — значение признака x_j для i-й единицы;
$\bar{x}_j$ — среднее значение признака x_j.

Пример. Рассмотрим использование многомерных средних на сельскохозяйственных предприятиях Всеволожского района Ленинградской области за 1999 г. (табл. 6.9). По каждому предприятию приведены четыре признака:

- среднемесячная оплата труда работника за определяемый вид работ, руб., x_1;
- валовой доход на 1 га сельскохозяйственных угодий, тыс. руб./га, x_2;
- среднегодовая стоимость основных производственных фондов на 1 га сельскохозяйственных угодий, тыс. руб./га, x_3;
- отношение дебиторской задолженности к кредиторской задолженности, %, x_4.

Таблица 6.9

Характеристики предприятий Всеволожского района Ленинградской области в 1999 г.

Предприятия	Значения признаков				В % к средней				Многомерная средняя, %
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	
«Ручьи»	597	390	20,6	72	148	199	106	107	140
«Бугры»	353	96	12,1	30	88	49	62	45	61
«Пригородное»	403	84	20,6	26	100	43	106	39	72
«Авлога»	231	71	15,1	74	57	36	78	110	70
«Всеволожское»	330	114	14,8	159	82	58	76	237	113
«Выборгское»	540	235	24,0	26	134	120	184	39	104
«Приневское»	372	461	33,2	85	93	235	171	127	156
«Щеглово»	393	113	15,0	62	98	58	77	92	81
Средние величины	402	196	19,4	67	100	100	100	100	100
Средние квадратические отклонения	109	142	6,4	41	—	—	—	—	—

Эти признаки можно считать однородными, так как бóльшая их величина положительно характеризует экономику предприятия. Предпочтительнее обобщать в многомерной средней признаки, либо все «положительные», либо все «отрицательные» (чем больше, тем хуже).

Многомерные средние, приведенные в последней графе табл. 6.9, обобщают четыре признака. При этом значимость признаков для оценки предприятия полагается одинаковой, что, конечно, спорно. Можно усложнить методику, приписав признакам на основе экспертной оценки разные веса, и вычислить взвешенные многомерные средние.

Судя по полученным значениям $\bar{p}_i$, предприятия делятся на группы с многомерными средними ниже 100% (четыре предприятия), несколько выше 100% (два предприятия) и резко превышающие 100% (два предприятия).

При большом объеме совокупности для выделения групп на основе многомерной средней необходимо установить интервалы значений многомерной средней:

$$i_{\bar{p}} = \frac{\bar{p}_{\max} - \bar{p}_{\min}}{\text{число групп}}.$$

Затем следует провести группировку единиц: определить их количество в каждой группе и постараться указать, в чем состоят качественные различия между группами.

Более обоснованным методом многомерной классификации является *кластерный анализ*. Само название метода этимологически берет начало от слов «класс», «классификация». Английское слово «the cluster» имеет значения: группа, пучок, куст, т.е. объединение каких-то однородных явлений. В данном контексте оно близко к математическому понятию «множество», причем, как и множество, кластер может содержать только одно явление, но не может в отличие от множества быть пустым.

Каждая единица совокупности в кластерном анализе рассматривается как точка в заданном признаковом пространстве. Значение каждого из признаков у данной единицы служит ее координатой в этом пространстве по аналогии с координатами точки в нашем реальном трехмерном пространстве. Таким образом, *признаковое пространство* — это область варьирования всех признаков совокупности изучаемых явлений. Если мы уподобим это пространство обычному пространству, имеющему евклидову метрику, то тем самым получим возможность измерять «расстояния» между точками признакового пространства. Эти расстояния называют *евклидовыми*. Их вычисляют по тем же правилам, что и в обычной евклидовой геометрии. На плоскости, т.е. в двухмерном пространстве, расстояние между точками A и B равно корню квадратному из суммы квадратов разностей координат этих точек по оси абсцисс и по оси ординат — на основе теоремы Пифагора (рис. 6.1):

$$d = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

В многомерном признаковом пространстве расстояние между точками p и q с k координатами, т.е. индивидуальными значениями k признаков, определяется так:

$$d_{p,\,q} = \sqrt{\sum_{j=1}^{k} (x_{jp} - x_{jq})^2}. \qquad (6.14)$$

Совершенно очевидно, что нельзя суммировать квадраты отклонений одной точки от другой в абсолютных значениях

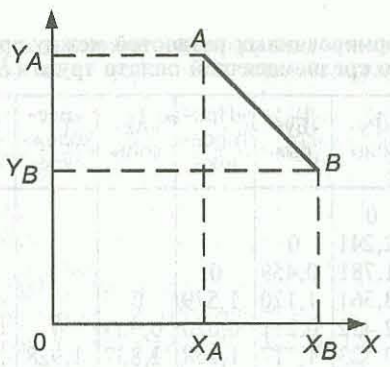


Рис. 6.1. Евклидово расстояние

разнокачественных признаков. Необходимо сначала выразить различия между единицами совокупности по каждому признаку в каком-то относительно безразмерном показателе. В качестве такого показателя часто применяют «нормированную разность», т.е. величину:

$$d_{j_{p,\,q}} = \frac{x_{jp} - x_{jq}}{\sigma_{xj}}, \qquad (6.15)$$

где $x_{jp} - x_{jq}$ — абсолютная разность значений j-го признака у единиц совокупности с номерами p и q;

σ_{xj} — среднее квадратическое отклонение признака x_j.

Знаки нормированных разностей не имеют значения, так как «расстояние» в признаковом пространстве — скалярная, а не векторная величина.

Пример. По данным табл. 6.9 среднее квадратическое отклонение признака x_1 равно 109[1]. Разделив все попарные раз-

[1] Здесь и далее величина среднего квадратического отклонения получена по формуле $\sigma_j = \sqrt{\sum_{(i)} (x_{ij} - \bar{x}_j)^2 / n}$. Поскольку рассматриваемые хозяйства — это выборка из всех сельскохозяйственных предприятий, то эта оценка дисперсии — смещенная. Правильнее было бы использовать несмещенную оценку дисперсии $s_j^2 = \frac{\sum (x_{ij} - \bar{x}_j)^2}{n-1}$, тогда среднее квадратическое отклонение для x_1 было бы равно 116, а не 109.

Таблица 6.10

Матрица нормированных разностей между предприятиями по среднемесячной оплате труда (D_1)

Предприятие	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	2,241	0						
«Пригородное»	1,781	0,459	0					
«Авлога»	3,361	1,120	1,579	0				
«Всеволожское»	2,452	0,211	0,670	0,909	0			
«Выборгское»	0,523	1,717	1,258	2,837	1,928	0		
«Приневское»	2,066	0,174	0,285	1,295	0,386	1,543	0	
«Щеглово»	1,873	0,367	0,092	1,488	0,579	1,350	0,193	0

Средняя нормированная разность $\overline{d}_1 = 1{,}182$.

ности значений этого признака на 109, получим матрицу нормированных разностей D_1 (табл. 6.10). Очевидно, эта матрица размера $n \times n$ симметрична, поскольку расстояние между A и B равно расстоянию между B и A.

Из данных табл. 6.10 видно, что величина нормированных разностей по этому признаку варьирует от 0 до 3,4. В нормально распределенной совокупности различия признака в среднем лишь в трех случаях из тысячи превосходят шесть сигм, т.е. в распределениях, близких к нормальным, величина нормированного расстояния редко превосходит 6.

В нормально распределенной совокупности $\overline{d}_{q,\,p}$ совпадает со средним отклонением нормированных разностей от средней величины, т.е. нормированная разность в нормальной совокупности в среднем равна единице, $\overline{d}_{q,\,p} = 1$. Это очень важно при установлении предельного (критического) расстояния в признаковом пространстве, при достижении которого прекращается объединение кластеров.

Аналогично вычисляются матрицы нормированных разностей по признакам x_2, x_3, x_4 (табл. 6.11—6.13).

На основе данных табл. 6.10—6.13 формируется матрица евклидовых расстояний D (табл. 6.14).

С учетом нормировки разности признаков расстояние между двумя любыми единицами совокупности (точками в признаковом пространстве) имеет вид:

Таблица 6.11

Матрица нормированных разностей между предприятиями по валовому доходу на 1 га сельскохозяйственных угодий (D_2)

Предприятие	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	2,070	0						
«Пригородное»	2,156	0,085	0					
«Авлога»	2,248	0,176	0,092	0				
«Всеволожское»	1,945	0,127	0,211	0,303	0			
«Выборгское»	1,092	0,979	1,064	1,155	0,853	0		
«Приневское»	0,500	2,572	2,656	2,748	2,445	1,592	0	
«Щеглово»	1,952	0,120	0,204	0,296	0,007	0,860	2,452	0

Средняя нормированная разность $\overline{d}_2 = 1{,}177$.

$$d_{p,\,q} = \sqrt{\sum_{j=1}^{k} d_{jp,\,q}^2}. \qquad (6.16)$$

Например, расстояние между предприятиями «Ручьи» и «Бугры» согласно формуле (6.16), по данным табл. 6.10—6.13, составляет:

$$d_{1,2} = \sqrt{2{,}241^2 + 2{,}071^2 + 1{,}330^2 + 1{,}018^2} = 3{,}481 \text{ и т.д.}$$

Таблица 6.12

Матрица нормированных разностей между предприятиями по среднегодовой стоимости основных производственных фондов на 1 га сельскохозяйственных угодий (D_3)

Предприятие	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	1,330	0						
«Пригородное»	0,000	1,330	0					
«Авлога»	0,861	0,469	0,861	0				
«Всеволожское»	0,908	0,422	0,908	0,047	0			
«Выборгское»	0,532	1,862	0,532	1,393	1,439	0		
«Приневское»	1,971	3,301	1,971	2,832	2,879	1,439	0	
«Щеглово»	0,876	0,454	0,876	0,016	0,031	1,408	2,848	0

Средняя нормированная разность $\overline{d}_3 = 1{,}207$.

Таблица 6.13

Матрица нормированных разностей между предприятиями по отношению дебиторской задолженности к кредиторской (D_4)

Предприятие	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	1,018	0						
«Пригородное»	1,115	0,097	0					
«Авлога»	0,048	1,066	1,163	0				
«Всеволожское»	2,109	3,127	3,224	2,060	0			
«Выборгское»	1,115	0,097	0,000	1,163	3,224	0		
«Приневское»	0,315	1,330	1,430	0,267	1,794	1,430	0	
«Щеглово»	0,242	0,776	0,873	0,291	2,351	0,873	0,557	0

Средняя нормированная разность $\overline{d}_4 = 1{,}184$.

Матрица евклидовых расстояний D служит основой *агломеративно-иерархического метода классификации*, который заключается в последовательном объединении группируемых объектов — сначала самых близких, а затем все более удаленных друг от друга. Процедура классификации состоит из последовательных шагов, на каждом из которых проводится объединение двух ближайших групп объектов (кластеров). На нулевом шаге каждый объект рассматривается как отдельный кластер.

Таблица 6.14

Матрица нормализованных (нормированных) евклидовых расстояний (D)

Предприятие	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	3,481	0						
«Пригородное»	3,011	1,413	0					
«Авлога»	4,134	1,626	2,144	0				
«Всеволожское»	3,881	3,165	3,422	2,273	0			
«Выборгское»	1,730	2,717	1,731	3,561	4,112	0		
«Приневское»	2,916	4,395	3,615	4,161	4,199	3,005	0	
«Щеглово»	2,854	0,978	1,257	1,544	2,421	2,303	3,804	0

Таблица 6.15

Нормированные разности и евклидовы расстояния для кластера «Бугры + Щеглово»

Предприятие	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние величины по кластеру	373	104,5	13,55	46	0
«Ручьи»	2,057	2,011	1,103	0,630	3,145
«Пригородное»	0,275	0,144	1,103	0,485	1,244
«Авлога»	1,304	0,236	0,243	0,679	1,509
«Всеволожское»	0,395	0,067	0,196	2,739	2,775
«Выборгское»	1,533	0,919	1,635	0,485	2,470
«Приневское»	0,009	2,512	3,075	0,945	4,082

На первом шаге объединим в кластер предприятия с наименьшим евклидовым расстоянием («Бугры» и «Щеглово»). Найдем средние по всем признакам для этого кластера и евклидовы расстояния от кластера до других предприятий (табл. 6.15).

Заменив в матрице евклидовых расстояний (см. табл. 6.14) расстояния предприятий, вошедших в первый кластер, на числа последней графы табл. 6.15, видим, что теперь минимальным является расстояние между предприятием «Пригородное» и первым кластером: $d = 1{,}244$ (табл. 6.16).

Таблица 6.16

Матрица евклидовых расстояний после образования кластера «Бугры + Щеглово» (Б + Щ)

Предприятие	Б + Щ	«Ручьи»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»
Кластер «Б + Щ»	0						
«Ручьи»	3,145	0					
«Пригородное»	1,244	3,011	0				
«Авлога»	1,509	4,134	2,144	0			
«Всеволожское»	2,775	3,881	3,422	2,273	0		
«Выборгское»	2,470	1,730	1,731	3,561	4,112	0	
«Приневское»	4,082	2,916	3,615	4,161	4,199	3,005	0

Таблица 6.17

Нормированные разности и евклидовы расстояния для кластера «Бугры + Щеглово + Пригородное»

Предприятие	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние величины по кластеру	383	97,7	15,9	39,3	0
«Ручьи»	1,965	2,060	0,735	0,792	3,045
«Авлога»	1,396	0,188	0,125	0,840	1,645
«Всеволожское»	0,487	0,115	0,172	2,900	2,948
«Выборгское»	1,442	0,968	1,267	0,323	2,174
«Приневское»	0,101	2,560	2,707	1,107	3,888

Таблица 6.18

Матрица евклидовых расстояний после образования кластера «Бугры + Щеглово + Пригородное» (Б + Щ + П)

Предприятие	Кластер «Б + Щ + + П»	«Ручьи»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»
Кластер «Б + + Щ + П»	0					
«Ручьи»	3,045	0				
«Авлога»	1,645	4,134	0			
«Всеволожское»	2,948	3,881	2,273	0		
«Выборгское»	2,174	1,730	3,561	4,112	0	
«Приневское»	3,888	2,916	4,161	4,199	3,005	0

Таблица 6.19

Нормированные разности и евклидовы расстояния для кластера «Бугры + Щеглово + Пригородное + Авлога»

Предприятие	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние величины по кластеру	345	91	15,7	48	0
«Ручьи»	2,312	2,106	0,766	0,585	3,273
«Всеволожское»	0,138	0,162	0,141	2,707	2,719
«Выборгское»	1,789	1,014	1,297	0,537	2,490
«Приневское»	0,248	2,606	2,734	0,902	3,891

Таблица 6.20

Матрица евклидовых расстояний после образования кластера «Бугры + Щеглово + Пригородное + Авлога» («Б + Щ + П + А»)

Предприятие	Кластер «Б + + Щ + П + А»	«Ручьи»	«Всеволожское»	«Выборгское»	«Приневское»
Кластер «Б + + Щ + П + А»	0				
«Ручьи»	3,273	0			
«Всеволожское»	2,719	3,881	0		
«Выборгское»	2,490	1,730	4,112	0	
«Приневское»	3,891	2,916	4,199	3,005	0

Таблица 6.21

Нормированные разности и евклидовы расстояния для кластеров 1 и 2

Предприятие	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние величины по кластеру 2	568,5	312,5	22,3	49	0
Кластер 1	2,052	1,561	1,033	0,024	2,778
«Всеволожское»	2,190	1,399	1,173	2,666	3,903
«Приневское»	1,804	1,046	1,705	0,873	2,832

Таблица 6.22

Матрица евклидовых расстояний после образования кластера 2

	Кластер 1	Кластер 2	Кластер 3 «Всеволожское»	Кластер 4 «Приневское»
Кластер 1	0			
Кластер 2	2,778	0		
Кластер 3 «Всеволожское»	2,719	3,903	0	
Кластер 4 «Приневское»	3,891	2,832	4,199	0

Следовательно, на втором шаге к первому кластеру присоединяется предприятие «Пригородное». Вычисляем средние величины, нормированные разности по каждому признаку и евклидовы расстояния от кластера, включающего три предприятия («Бугры», «Щеглово», «Пригородное»), до каждого из оставшихся предприятий. Результаты представлены в табл. 6.17.

Заменив евклидовы расстояния предприятий, вошедших в кластер, данными последней графы табл. 6.17, получим новую матрицу евклидовых расстояний (табл. 6.18).

Минимальным является евклидово расстояние от кластера до предприятия «Авлога». На третьем шаге образуем кластер «Бугры + Щеглово + Пригородное + Авлога». Полученные средние величины для кластера, нормированные разности и евклидовы расстояния представлены в табл. 6.19 и 6.20.

Минимальное евклидово расстояние между предприятиями «Ручьи» и «Выборгское» меньше двух, следовательно, эти предприятия объединяются в кластер 2 (табл. 6.21). Кластер «Б + Щ + П + А» будем называть кластером 1.

После четвертого шага получаем новую матрицу евклидовых расстояний (табл. 6.22).

Согласно табл. 6.22 все расстояния больше двух. Оставляем четыре типа предприятий: предприятия, вошедшие в кластер 1, кластер 2, кластер 3 («Всеволожское») и кластер 4 («Приневское»).

Сравнивая результат кластерного анализа с многомерными средними (см. табл. 6.9), видим, что состав кластера 1 точно отвечает тем хозяйствам, чьи многомерные средние ниже 100%. Также выделение в самостоятельный кластер предприятия «Приневское» соответствует его высшему значению многомерной средней. А вот объединение в кластер 2 предприятий «Ручьи» и «Выборгское» не соответствует многомерным средним, по которым к предприятию «Ручьи» было ближе предприятие «Всеволожское». В результате резкого отличия по признаку x_4 предприятие «Всеволожское» выделилось в отдельный кластер 3.

Обобщая рассмотренную процедуру кластерного анализа, представим действия в виде определенной последовательности.

1. Вычисление средних величин для каждого из классификационных признаков $\bar{x}_j$ в целом по совокупности.

2. Вычисление средних квадратических отклонений для каждого из признаков по совокупности σ_{xj} или s_{xj}.

3. Вычисление матрицы нормированных разностей по каждому из группировочных признаков — $d_{jp,\,q}$.

4. Вычисление евклидовых расстояний между каждой парой сочетаний единиц совокупности — $d_{p,\,q}$.

5. Выбор наименьшего из евклидовых расстояний — $\min d_{p,\,q}$.

6. Объединение единиц совокупности с наименьшим евклидовым расстоянием между ними в один кластер.

7. Вычисление средних значений всех признаков для единиц, объединенных в кластер.

8. Вычисление новых нормированных расстояний между объединенным кластером и остальными единицами.

9. Вычисление новых евклидовых расстояний между объединенным кластером и остальными единицами (или кластерами).

10. Выбор наименьшего из евклидовых расстояний.

11. Повторение операций (пункты 6—10) и т.д.

Объединение в кластеры прекращается, когда все евклидовы расстояния превысят заданную критическую величину $d_{\text{крит}}$. Обычно ППП предусматривает вывод на печать состава (перечня единиц совокупности) каждого кластера, евклидовых расстояний между ними, матриц нормированных разностей по каждому признаку.

Существует много достаточно сложных алгоритмов кластерного анализа и родственных ему методов распознавания образов, таксономии и др.

Рассмотренная выше методика вычисления евклидова расстояния предполагает, что все признаки считаются равноправными. На самом же деле при выделении типов социально-экономических явлений группировочные признаки не равноправны: как правило, одни признаки имеют большее, другие — меньшее значение. Следовательно, более совершенная методика кластерного анализа должна учитывать разную значимость, разный «вес» группировочных признаков. В этом случае должно использоваться *взвешенное евклидово расстояние*:

$$d_{p,\,q} = \sqrt{\sum_{j=1}^{k} d_{jp,\,q} w_j}, \qquad (6.17)$$

где w_j — вес j-го признака.

Определение весов — весьма сложная задача, выходящая за пределы компетенции статистики. О том, какие признаки важнее при классификации тех или иных объектов, могут судить не статистики, а специалисты в соответствующей отрасли. Поэтому одним из способов определения весов признаков при кластерном анализе являются оценки экспертов. Опросив специалистов-экспертов (не менее 6—10), статистик сможет определить по их оценкам место (роль) каждого группировочного признака. Затем найти средний «вес» признака. Можно просить экспертов ранжировать признаки по порядку значимости и определять «среднее место», но оценка при этом будет очень грубая: признак, поставленный на первое место, будет вдвое важнее второго и в двадцать или тридцать раз важнее последнего. Для того чтобы различия весов были не столь значительными, можно просить экспертов распределить общую сумму оценок (100 или 1000%) между группировочными признаками в соответствии с их значениями. Тогда каждому из признаков будет приписана некоторая доля этой общей суммы, можно двум-трем признакам приписать одинаковые веса. Но этот способ взвешивания требует от экспертов большей точности и напряжения, чем простое ранжирование признаков.

Субъективность экспертных оценок в какой-то мере можно компенсировать статистической обработкой. Например, по каждому признаку перед определением средней оценки его веса можно отбросить максимальную и минимальную оценки, если они существенно отличаются от оценок остальных экспертов. Можно вообще исключить того эксперта, чьи оценки в среднем отличаются от средних оценок признаков более чем, например, на 2σ. Однако эти статистические коррективы небезупречны и допустимы при значительном числе экспертов для того, чтобы их средние оценки были надежны.

Существует и другая возможность оценки роли группировочных признаков, их значимости для классификации: на основе стандартизованных коэффициентов регрессии или коэффициентов раздельной детерминации (гл. 9).

Рассмотренный алгоритм иерархической классификации можно модифицировать, используя *метод «ближайшего» или «дальнего соседа»* (табл. 6.23). В этом случае в матрицу евклидовых расстояний вводятся расстояния, полученные не на основе средних величин по кластеру; в качестве представителя

кластера берется входящий в него объект либо наименее удаленный от остальных («ближайший сосед»), либо наиболее удаленный от остальных («дальний сосед»). Поскольку min $d_{8,2} = 0{,}978$ (см. табл. 6.14), предприятия «Бугры» и «Щеглово» были объединены в кластер. При использовании метода «ближайшего соседа» в последующей после объединения этих двух предприятий матрице евклидовых расстояний кластер будет представлять то «Бугры», то «Щеглово» — в зависимости от того, какое из предприятий наименее удалено от остальных. Для простоты будем использовать не названия, а порядковые номера предприятий, соответствующие их последовательности в табл. 6.9.

Таблица 6.23

Матрица евклидовых расстояний на первом шаге (метод «ближайшего соседа»)

№ п/п	Предприятие	1	3	4	5	6	7	8 + 2
1	«Ручьи»	0						
3	«Пригородное»	3,011	0					
4	«Авлога»	4,134	2,144	0				
5	«Всеволожское»	3,881	3,442	2,273	0			
6	«Выборгское»	1,730	1,731	3,561	4,112	0		
7	«Приневское»	2,916	3,615	4,161	4,199	3,005	0	
8 + 2	«Бугры + + Щеглово»	2,854	1,257	1,544	2,421	2,303	3,804	0

Минимальное евклидово расстояние между кластером и предприятием «Пригородное» $d_{8,\,2,\,3} = 1{,}257$ (см. табл. 6.14). Это хозяйство имеет номер 3, присоединим его к кластеру (8 + 2). Матрица евклидовых расстояний на втором шаге будет следующей (табл. 6.24).

Минимальным является расстояние между предприятием 4 («Авлога») и кластером (в качестве ближайшего к «Авлоге» соседа выступает хозяйство «Щеглово»): min $d_{8,\,2,\,3,\,4} = 1{,}544$. Матрица евклидовых расстояний при объединении четырех предприятий в кластер представлена в табл. 6.25.

Минимальным в табл. 6.25 оказалось расстояние между хозяйствами 1 и 6 («Ручьи» и «Выборгское»), которые на этом шаге образуют кластер (табл. 6.26).

В табл. 6.26 между двумя выделенным кластерами определяется минимальное расстояние, а именно — дистанция ме-

жду хозяйствами 3 и 6 («Пригородное» и «Выборгское»): min $d_{8,2,3,4,6} = 1{,}730$. Поскольку min $d < 2$, то хозяйства 1 и 6 присоединяются к первому кластеру (табл. 6.27).

Расстояние между кластером и хозяйством 5, так же как и на предыдущем шаге, определяется минимальным расстоянием. В данном случае это расстояние между хозяйствами 5 и 4. Расстояние между предприятием 7 и кластером стало определяться «ближайшим соседом» — предприятием 1. Все расстояния в табл. 6.27 $d_{p,q} > 2$.

Таблица 6.24

Матрица евклидовых расстояний на втором шаге (метод «ближайшего соседа»)

Предприятие	1	4	5	6	7	8 + 2 + 3
1	0					
4	4,134	0				
5	3,881	2,273	0			
6	1,730	3,561	4,112	0		
7	2,916	4,161	4,199	3,005	0	
8 + 2 + 3	2,854	1,544	2,421	1,730	3,615	0

Таблица 6.25

Матрица евклидовых расстояний на третьем шаге (метод «ближайшего соседа»)

Предприятие	1	5	6	7	8 + 2 + 3 + 4
1	0				
5	3,881	0			
6	1,730	4,112	0		
7	2,916	4,199	3,005	0	
8 + 2 + 3 + 4	2,854	2,273	3,615	1,731	0

Таблица 6.26

Матрица евклидовых расстояний на четвертом шаге (метод «ближайшего соседа»)

Предприятие	1 + 6	8 + 2 + 3 + 4	5	7
1 + 6	0			
8 + 2 + 3 + 4	1,730	0		
5	3,881	2,273	0	
7	2,916	3,615	4,199	0

Если придерживаться некоторого критического значения евклидова расстояния $d_{крит} = 2$, как в ранее рассмотренном примере, то в итоге совокупность подразделяется на три кластера, два из которых содержат по одному предприятию (5 и 7) и один — шесть предприятий.

Представим процесс классификации в виде дендрограммы (рис. 6.2).

Дендрограмма — дерево объединений кластеров с порядковыми номерами объектов по горизонтальной оси и шкалой расстояний по вертикальной оси.

Решение, полученное методом «ближайшего соседа», близко к прежнему результату при описании кластеров средними показателями (см. табл. 6.22), но не совпадает с ним — вместо четырех кластеров здесь выделились три.

Если применить метод «дальнего соседа», то на первом шаге, после объединения предприятий 2 и 8, получим следующую матрицу евклидовых расстояний (табл. 6.28).

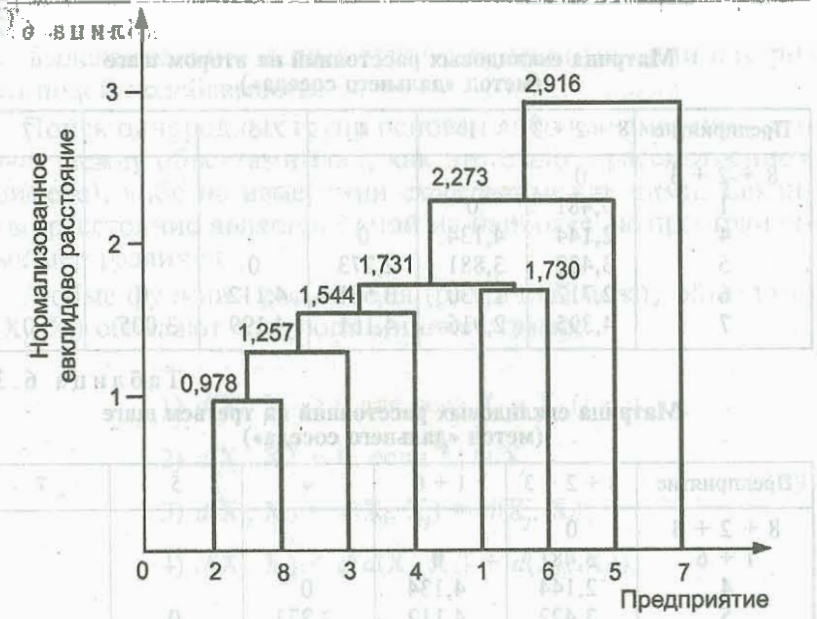


Рис. 6.2. Дендрограмма: метод «ближайшего соседа»

Таблица 6.27

Матрица евклидовых расстояний на пятом шаге (метод «ближайшего соседа»)

Предприятие	8 + 2 + 3 + 4 + 1 + 6	5	7
8 + 2 + 3 + 4 + 1 + 6	0		
5	2,273	0	
7	2,916	4,199	0

Таблица 6.28

Матрица евклидовых расстояний на первом шаге (метод «дальнего соседа»)

Предприятие	1	3	4	5	6	7	8 + 2
1	0						
3	3,011	0					
4	4,134	2,144	0				
5	3,881	3,422	2,273	0			
6	1,730	1,731	3,561	4,112	0		
7	2,916	3,615	4,161	4,199	3,005	0	
8 + 2	3,481	1,431	1,626	3,165	2,717	4,395	0

Таблица 6.29

Матрица евклидовых расстояний на втором шаге (метод «дальнего соседа»)

Предприятие	8 + 2 + 3	1	4	5	6	7
8 + 2 + 3	0					
1	3,481	0				
4	2,144	4,134	0			
5	3,422	3,881	2,273	0		
6	2,717	1,730	3,561	4,112	0	
7	4,395	2,916	4,161	4,199	3,005	0

Таблица 6.30

Матрица евклидовых расстояний на третьем шаге (метод «дальнего соседа»)

Предприятие	8 + 2 + 3	1 + 6	4	5	7
8 + 2 + 3	0				
1 + 6	3,481	0			
4	2,144	4,134	0		
5	3,422	4,112	2,273	0	
7	4,395	3,005	4,161	4,199	0

Табл. 6.28 отличается от табл. 6.21 последней строкой, в которой показаны максимальные расстояния кластера (8 + 2) от других объектов.

На втором шаге выбирается наименьшее из $d_{p,q}$. В данном примере это расстояние между хозяйством 3 и кластером (8 + + 2): $d_{8,2;3} = 1{,}431$. Образуется новый кластер (8 + 2 + 3), в котором также выделяется «дальний сосед» (табл. 6.29).

В табл. 6.29 $d_{\min} = d_{1,6} = 1{,}730$. Таким образом возникает второй кластер (1 + 6), описанный в табл. 6.30.

В табл. 6.30 все значения $d_{p,q} > 2$. Следовательно, в результате метода «дальнего соседа» получаем пять кластеров, три из которых включают по одному предприятию. В определенном смысле это худший вариант классификации.

Подведем итоги.

Все алгоритмы многомерной классификации основаны на целевой функции:

$$\sum_{1}^{m}\sum_{1}^{n_i}(x_{ij}-\bar{x}_j)^2 \to \min, \qquad (6.18)$$

т.е. выделение однородных групп при минимизации внутригрупповой колеблемости.

Поиск однородных групп основан либо на измерении *различия* между объектами (так, как это было в рассмотренном примере), либо на измерении *сходства* между ними. Евклидово расстояние является одной из наиболее распространенных мер различия.

Любые функции расстояния (различия) между объектами $d(\mathbf{X}_i, \mathbf{X}_j)$ обладают следующими свойствами:

$$\begin{aligned}
&1)\ d(\mathbf{X}_i, \mathbf{X}_j) > 0 \text{ для всех } \mathbf{X}_i \text{ и } \mathbf{X}_j\ (i \neq j);\\
&2)\ d(\mathbf{X}_i, \mathbf{X}_j) = 0, \text{ если } \mathbf{X}_i = \mathbf{X}_j;\\
&3)\ d(\mathbf{X}_i, \mathbf{X}_j) = d(\mathbf{X}_i, \mathbf{X}_j) = d(\mathbf{X}_j, \mathbf{X}_i);\\
&4)\ d(\mathbf{X}_i, \mathbf{X}_j) < d(d(\mathbf{X}_i, \mathbf{X}_h) + d(\mathbf{X}_h, \mathbf{X}_j)),
\end{aligned} \qquad (6.19)$$

где $\mathbf{X}_i$, $\mathbf{X}_j$, $\mathbf{X}_h$ — любые три вектора.

Расстояния между парами векторов $d(\mathbf{X}_i, \mathbf{X}_j)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}. \qquad (6.20)$$

Диагональные элементы d_{ii} для всех i равны нулю.

Расстояние между кластером $i + j$ и всеми другими кластерами вычисляется в соответствии с выбранной стратегией классификации как:

«ближайшего соседа»: $d_{i+j,\,k} = \min(d_{ik}, d_{jk})$;

«дальнего соседа»: $d_{i+j,\,k} = \max(d_{ik}, d_{jk})$;

«группового соседа»: $d_{i+j,\,k} = (d_{ik}n_i + d_{jk}n_j)/(n_i + n_j)$,

где $n_i + n_j = n_{i+j}$.

Метод «ближайшего соседа» сжимает пространство исходных переменных и рекомендуется для получения минимального дерева иерархической классификации. Метод «дальнего соседа» растягивает пространство. Метод «группового соседа» сохраняет метрику пространства.

Если классификация данных основана на мерах сходства $s(\mathbf{X}_i, \mathbf{X}_j)$, то следует иметь в виду общие свойства этих мер:

$$\begin{aligned} &1)\ 0 \leqslant s(\mathbf{X}_i, \mathbf{X}_j) < 1 \text{ для } \mathbf{X}_i \neq \mathbf{X}_j; \\ &2)\ s(\mathbf{X}_i, \mathbf{X}_i) = 1; \\ &3)\ s(\mathbf{X}_i, \mathbf{X}_j) = s(\mathbf{X}_j, \mathbf{X}_i). \end{aligned} \qquad (6.21)$$

Соответственно матрица мер сходства имеет вид:

$$S = \begin{pmatrix} 1 & s_{12} & \dots & s_{1n} \\ s_{21} & 1 & \dots & s_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ s_{n1} & s_{n2} & \dots & 1 \end{pmatrix}. \qquad (6.22)$$

Диагональные элементы такой матрицы равны единице.

В качестве мер сходства чаще всего используются коэффициенты корреляции (гл. 9).

Основными ППП для решения задачи многомерной классификации являются «Класс-мастер», SPSS, SAS. Как правило, алгоритмы неиерархической многомерной классификации основаны на геометрическом представлении кластера как локального скопления точек в заданном признаковом пространстве.

Большинство методов классификации основано на однозначном отнесении объекта к тому или иному классу. Но, как уже отмечалось, границы классов могут быть размытыми, нечеткими. Класс объектов, в котором нет резкой границы между объектами, входящими в него, и теми, которые в него не входят, называется *нечетким множеством*.

Для классификации данных в нечетких множествах необходимо ввести матрицу принадлежности каждого объекта к нечеткому множеству с элементами

$$\mu_{ij} \in [0, 1]\ \forall\, i \in 1 \div n,\ j \in 1 \div k.$$

Для μ_{ij} выполняется следующее условие: $\sum_{j=1}^{k} \mu_{ij} = 1$, т.е. объект обязательно принадлежит к тому или иному нечеткому множеству. Качество разбиения определяется как минимизацией внутриклассовой дисперсии, так и максимизацией удаленности центров классов.

Алгоритмы и программы многомерной классификации постоянно развиваются: разрабатываются ППП, учитывающие размытость границ между классами (распознавание в нечетких множествах), различную длину описаний классов и т.д. Большое значение в решении задач иерархических классификаций имеет компьютерная графика — так называемые классификационные деревья. Подробнее вопросы многомерной классификации освещаются в работах, указанных в списке рекомендуемой литературы.

РЕЗЮМЕ

Требование однородности данных выдвигается на всех этапах статистического анализа. Для получения однородных данных проводится группировка. При этом различия между единицами, отнесенными к одной группе, должны быть меньше, чем между единицами, отнесенными к разным группам.

Проведение группировки включает выбор группировочного признака (или признаков) и определение границ интервалов.

Чаще всего группировки проводятся с равными интервалами, но при неравномерном изменении группировочного признака и его значительной вариации применяются группировки с равнонаполненными интервалами.

В зависимости от цели проведения различают следующие виды группировок: типологические, структурные, аналитические.

Типологическая группировка проводится с целью выделения социально-экономических типов.

Структурная группировка соответствует вариационному ряду.

Аналитическая группировка строится для изучения зависимости одного признака от другого. На ее основе измеряются сила и теснота связи, т.е. вычисляется эмпирическое корреляционное отношение. Для погашения влияния прочих факторов в аналитической группировке целесообразно рассчитывать стандартизованные групповые средние. Выводы о характере и интенсивности связи между признаками во многом зависят от выбранного числа групп.

При необходимости группировки по многим признакам для каждой единицы рассчитывают многомерную среднюю, а затем по ее значениям группируют данные.

Многомерные группировки часто называют многомерными классификациями. Они бывают иерархические, неиерархические, основанные на мерах различия или сходства. В качестве меры различия чаще всего используется евклидово расстояние. Среди иерархических классификаций выделяются метод средних, метод «ближайшего соседа», метод «дальнего соседа».

Исходя из структуры типа (ядро + слой) развиваются вероятностные классификации, так называемые классификации в размытых (нечетких) множествах.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Айвазян С. А., Мхитарян В. С.* Теория вероятностей и прикладная статистика. Т. 1: Учебник для вузов. — 2-е изд. — М.: ЮНИТИ, 2001.
2. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ: Пер. с англ. — М.: Мир, 1982.
3. *Елисеева И. И., Рукавишников В. О.* Группировка, корреляция, распознавание образов. — М.: Статистика, 1977.
4. *Енюков И. С.* Методы — алгоритмы — программы многомерного статистического анализа. — М.: Финансы и статистика, 1986.
5. *Козлов А. Ю., Шишлов В. Ф.* Пакет анализа MS Excel в экономико-статистических расчетах / Под ред. В. С. Мхитаряна. — М.: ЮНИТИ, 2003.
6. *Кулаичев А. П.* Методы и средства анализа данных в среде Windows. Stadia 6.0. — М.: НПО «Информатика и компьютеры», 1996.
7. *Мандель И. Д.* Кластерный анализ. — М.: Финансы и статистика, 1988.
8. *Миркин Б. Г.* Группировки в социально-экономических исследованиях. — М.: Финансы и статистика, 1985.

Глава
7

ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ. ИСПЫТАНИЕ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

7.1. Причины применения выборочного наблюдения. Дескриптивная статистика и статистический вывод

В гл. 2 отмечалось, что статистика далеко не всегда оперирует данными сплошного наблюдения. Из всех видов несплошного наблюдения главным является выборочное наблюдение, так как только выборочный метод имеет статистико-математическое обоснование распространения данных, полученных по выборке, на всю совокупность.

Причин использования выборочного метода несколько.

Во-первых, как это ни парадоксально, это *повышение точности данных*: уменьшение числа единиц наблюдения в выборке резко снижает ошибки регистрации. Правда, за счет неполноты охвата единиц возникает ошибка репрезентативности, т.е. представительности выборочных данных. Но даже взятые вместе ошибка наблюдения для выборки плюс ошибка репрезентативности обеспечивают большую точность выборочных данных по сравнению с массовым сплошным наблюдением.

При ограниченном объеме работ можно привлечь более квалифицированных исполнителей (интервьюеров, счетчиков-регистраторов). Это положительно сказывается на качестве данных выборочного обследования.

Во-вторых, обращение к выборкам обеспечивает *экономию материальных, трудовых, финансовых ресурсов и времени*. Например, для составления баланса денежных доходов и расходов населения, для изучения денежного обращения, выявления дифференциации населения по уровню жизни, определения черты бедности и т.д. необходимы данные о бюджетах домохозяйств. Сбор этих данных осуществляется государственной статистикой, но один статистик в состоянии курировать ежедневные записи доходов, расходов, потребления не более чем в 20—25 домохозяйствах. Если бы он решил собирать данные о бюджетах всех домохозяйств, то только для этой цели (не учитывая потребности последующей обработки) потребовалось бы примерно два миллиона статистиков. Так что использование выборочного наблюдения является единственным экономически выгодным решением, тем более что по результатам изучения сравнительно небольшой части можно получить с достаточно высокой степенью уверенности данные о всей совокупности. Подобная ситуация возникает при аудиторских проверках крупных фирм, когда вместо детального изучения каждого платежного документа ограничиваются анализом выборки документов, и в других областях применения статистики.

В-третьих, без выборки не обойтись, когда наблюдение связано с порчей наблюдаемых объектов. Это относится прежде всего к изучению качества продукции, которое основано на испытаниях образцов на вибрацию, упругость, разрыв и т.д. Всю продукцию, конечно же, таким испытаниям не подвергают, а только отобранные образцы. То же можно сказать об исследовании молока на жирность, зерна — на содержание белка, влажность, чистоту и всхожесть семян, электрических лампочек — на длительность горения и т.д. На выборках основаны маркетинговые исследования, оценки качества поставок.

Практика применения выборочного метода очень разнообразна. Иногда, проведя сплошное наблюдение, используют выборочный метод при *разработке данных*: отбирают часть данных для более подробной разработки по расширенной программе. Так поступают, например, при разработке данных переписи населения о составе и типах семей. Нередко в процессе сбора данных применяют *совместно сплошное и несплошное наблюдение*. При переписях населения в нашей стране (1959, 1970, 1979 гг.) собирались сведения о каждом лице по 11 признакам, а 25% населения давали более подробную информацию (18 вопросов).

Выборки используются при опросах общественного мнения, при выяснении потребительских предпочтений, формировании доходов и расходов населения, при определении урожайности сельскохозяйственных культур и продуктивности скота. С 20-х гг. XX в. выборочный метод стал использоваться для контроля и анализа качества продукции. Сейчас методы статистической выборки все шире внедряются в самые различные области. В 1994 г. в Российской Федерации была проведена 5%-ная микроперепись населения с целью уточнения демографического и социального состава населения, уровня благосостояния, включая жилищные условия, источники дохода и др. Эта микроперепись была положена в основу новой бюджетной выборки, созданной в 1996 г. на период до 2003 г., после чего она должна быть пересмотрена на основе данных Всероссийской переписи населения 2002 г.

Та совокупность, из которой проводится отбор, называется *генеральной совокупностью*; отобранные данные составляют *выборочную совокупность*. Эти данные представляют интерес, поскольку дают основание для суждений о параметрах и свойствах генеральной совокупности.

Таким образом, выборочный метод обладает следующими достоинствами:

- относительно небольшие (по сравнению со сплошным наблюдением) материальные, трудовые и стоимостные затраты на сбор данных (включая затраты на планирование и формирование выборки);
- оперативность получения результатов;
- широкая область применения;
- высокая достоверность результатов.

Все эти достоинства проявляются лишь при условии правильного решения проблем выборочного обследования. К ним относятся:

1) определение границ генеральной совокупности;

2) разработка программы наблюдения и инструкций;

3) определение основы для проведения выборки — списка единиц генеральной совокупности, сведений об их размещении и т.д.;

4) установление допустимого размера погрешности и определение объема выборки;

5) определение вида выборочного наблюдения;

6) установление сроков проведения наблюдения;

7) определение потребности в кадрах для проведения выборочного наблюдения, их подготовка;

8) оценка точности и достоверности данных выборки, определение порядка их распространения на генеральную совокупность.

Представление о статистических данных, как о выборочных, может относиться не только к собственно выборке, но и к данным сплошного наблюдения, которые иногда рассматриваются как выборка из всех возможных реализаций изучаемого процесса. Это имеет смысл в случае малого числа единиц совокупности. Кроме того, трактовка данных как выборочных используется применительно к результатам эксперимента, которые рассматриваются как некая выборка из потенциально бесконечного числа повторений экспериментальных наблюдений.

Трактовка данных как выборочных является основой деления статистики на *описательную* (*дескриптивную*) и *выводную*.

Методы описательной статистики включают сбор данных по всем единицам изучаемой совокупности, их обработку, получение сводных показателей, которые характеризуют только наблюдаемую совокупность. Например, если наша задача состоит в изучении успеваемости группы студентов, включающей 25 человек, то вычисленный средний балл по этой группе, процент отличных оценок и т.д. являются описаниями данной совокупности. Если же мы будем рассматривать эту группу студентов с точки зрения оценки успеваемости всех студентов данного колледжа или университета, то эта группа предстанет как выборка из общего числа студентов. В таком случае средний балл для группы будет являться *оценкой* средней успеваемости студентов колледжа в целом.

Генеральная совокупность может быть *реальной*, а может быть *гипотетической*, включающей случаи, которые реально не существуют, например, все возможные результаты эксперимента.

В выводной статистике принято строго различать *параметры* и *свойства* генеральной совокупности и их *оценки* по данным выборки. С этой целью принята следующая система обозначений: генеральные параметры обозначаются греческими буквами, выборочные показатели, которые рассматри-

ваются как оценки генеральных параметров, — латинскими буквами:

	Генеральные параметры	Выборочные показатели
Средняя величина	μ	$\bar{x}$
Относительная величина	π	p
Дисперсия	σ^2	s^2
Коэффициент корреляции	ρ	r

Объем генеральной совокупности обозначают N, объем выборочной совокупности — n.

Выборочные оценки отличаются от генеральных параметров за счет ошибки наблюдения и ошибки выборки:

Выборочная оценка = Генеральный параметр ± Ошибка наблюдения ± Ошибка выборки

Подводя итоги, можно сказать, что описательная статистика является инструментом описания совокупности, по которой у нас полностью имеются исходные данные.

Метод статистического вывода позволяет по данным выборок делать заключение о бóльшей совокупности, по которой мы не имеем исчерпывающих наблюдений.

7.2. Способы отбора, обеспечивающие репрезентативность выборки. Виды выборки

Для того чтобы по выборке можно было делать вывод о свойствах генеральной совокупности, выборка должна быть *репрезентативной* (представительной), т.е. полно и адекватно представлять свойства генеральной совокупности. Репрезентативность выборки может быть обеспечена только при *объективности отбора* данных.

Выборочная совокупность формируется по принципу массовых вероятностных процессов, без каких бы то ни было исключений из принятой схемы отбора. Необходимо обеспечить относительную однородность выборочной совокупности, или ее разделение на однородные группы единиц. При формировании выборочной совокупности должно быть дано четкое определение единицы отбора. Желателен приблизительно одинаковый размер единиц отбора, причем результаты будут тем точнее, чем меньше единица отбора.

Возможны три способа отбора: *случайный отбор, отбор единиц по определенной схеме, сочетание первого и второго способов.*

Если отбор в соответствии с принятым способом проводится из генеральной совокупности, предварительно разделенной на типы (слои или страты), то такая выборка называется *типической* (или расслоенной, или стратифицированной, или районированной). Еще одно деление выборки по видам определяется тем, что является единицей отбора: единица наблюдения или серия единиц (иногда используют термин «гнездо»). В последнем случае выборка называется *серийной* или *гнездовой*. На практике часто используется сочетание типической выборки с отбором сериями. В математической статистике, обсуждая проблему отбора данных, обязательно вводят деление выборки на *повторную* и *бесповторную*. Первая соответствует схеме возвратного шара, вторая — безвозвратного (при рассмотрении процесса отбора данных на примере отбора шаров разного цвета из урны). В социально-экономической статистике нет смысла применять повторную выборку, поэтому, как правило, имеется в виду бесповторный отбор. Если выборка проводится по схеме возвратного шара, то вероятность попадания любой единицы в выборку равна $1/N$, и она остается той же самой на протяжении всей процедуры отбора. Если выборка проводится по схеме невозвратного шара, то вероятность попадания единицы в выборку изменяется от $1/N$ — для первой отбираемой единицы, до $\frac{1}{N-n+1}$ — для последней.

Поскольку социально-экономические объекты имеют сложную структуру, организовать выборку бывает довольно трудно. Например, чтобы провести отбор домохозяйств при изучении потребления населения крупного города, легче провести сначала отбор территориальных ячеек, жилых домов, потом квартир или домохозяйств, затем респондента. Такая выборка называется *многоступенчатой*. На каждой ступени используются разные единицы отбора: более крупные — на начальных ступенях, на последней ступени единица отбора совпадает с единицей наблюдения.

Еще один вид выборочного наблюдения — *многофазовая выборка*. Такая выборка включает определенное количество

фаз, каждая из которых отличается подробностью программы наблюдения. Например, 25% всей генеральной совокупности обследуются по краткой программе, каждая четвертая единица из этой выборки обследуется по более полной программе и т.д.

При любом виде выборки отбор единиц проводится тремя отмеченными способами. Рассмотрим процедуру *случайного отбора*. Прежде всего составляется список единиц совокупности, в котором каждой единице присваивается цифровой код (номер или метка). Затем проводится жеребьевка. Шары с соответствующими номерами закладываются в барабан, перемешиваются, и проводится их отбор. Выпавшие номера соответствуют единицам, попавшим в выборку; число номеров равно запланированному объему выборки.

Отбор жеребьевкой может быть подвержен смещениям, вызванным недостатками техники (качеством шаров, барабана) и другими причинами. Более надежен с точки зрения объективности отбор по таблице случайных чисел. Такая таблица содержит серии цифр, чередующихся случайным образом, отобранных путем электронных сигналов. Поскольку мы пользуемся десятичной цифровой системой 0, 1, 2, ..., 9, вероятность появления любой цифры равна 1/10. Следовательно, если бы нужно было создать таблицу случайных чисел, включающую 500 знаков, то 50 из них были бы нули, столько же — единиц и т.д. Ввиду того, что каждая цифра и их последовательность являются случайными, можно использовать таблицу случайных чисел, перемещаясь либо по ее вертикали, либо по горизонтали. Цифры сгруппированы по пять для лучшей обозримости таблицы и пользования ею (табл. П.7 приложения).

Пример. Предположим, что нам нужно провести 5%-ную выборку из 9540 студентов университета. Объем выборки составит: $n = 5\% \cdot N = 477$ студентов.

Ввиду того, что объем генеральной совокупности выражается четырехзначным числом, код каждого студента должен быть четырехзначным: от 0001 — для первого студента до 9540 — для последнего студента в списке. Для того чтобы провести отбор по таблице случайных чисел, нужно выбрать начальную точку: можно закрыть глаза и поставить наугад точку в таблицу карандашом. Предположим, мы попали в 13-ю строку в 1-й столбец (табл. 7.1).

Таблица 7.1

Пример использования таблицы случайных чисел

Строка	Столбец							
	1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й
13-я	90822	60280	88925	99610	42772	60561	76873	04117
14-я	72121	79152	96591	90305	10189	79778	68016	13747
15-я	95268	41377	25684	08151	61816	58555	54305	86189
16-я	92603	09091	75884	93424	72586	88903	30061	14457
17-я	18813	90291	05275	01223	79607	85426	34900	09778
18-я	38840	26903	28624	67157	51986	42865	14508	49315

Следовательно, единица с номером 9082 является первой в выборке. Если двигаться по строке, то единица с номером 2602 будет второй, 8088 — третьей, 9259 — четвертой. Следующий код 9610 пропускаем, так как у нас нет студента с таким номером. Далее в выборку попадают номера 4277, 2605, 6176, 8730, 4117, 7212, 1791, 5296, 5919, 0305, 1018. Код 9797 пропускается. Следующие отобранные номера 7868, 0161, 3747, 9526, 8413, 7725 и т.д.

Процедура продолжается, пока число отобранных номеров не составит требуемый объем выборки ($n = 477$).

Часто используется отбор по какой-либо схеме (так называемая направленная выборка). Схема отбора принимается такой, чтобы отразить основные свойства и пропорции генеральной совокупности. Простейший способ — по спискам единиц генеральной совокупности, составленным так, чтобы упорядочивание единиц было бы не связано с изучаемыми свойствами, проводится *механический отбор* единиц с шагом, равным $N : n$. Обычно отбор начинают не с первой единицы, а отступив полшага, чтобы уменьшить возможность смещения выборки. Частота появления единиц с теми или иными особенностями, например студентов с тем или иным уровнем успеваемости, живущих в общежитии, и т.д., будет определяться той структурой, которая сложилась в генеральной совокупности.

Для большей уверенности в том, что выборка отразит структуру генеральной совокупности, последняя подразделяется на типы, и проводится случайный или механический от-

бор из каждого типа. Общее число единиц, отобранных из разных типов, должно соответствовать объему выборки.

Особые трудности возникают, когда нет списка единиц, а отбор нужно провести либо на местности, либо из образцов продукции на складе готовой продукции. В этих случаях важно детально разработать схему ориентации на местности и схему отбора и следовать ей, не допуская отклонений. Например, счетчик получает указание двигаться от определенной автобусной остановки на север по четной стороне улицы и, отсчитав два дома от первого угла, войти в третий и провести опрос в каждом пятом жилом помещении. Неукоснительное следование принятой схеме обеспечивает выполнение главного условия формирования репрезентативной выборки — объективность отбора единиц.

От случайной выборки следует отличать ***квотный отбор***, когда выборка конструируется из единиц определенных категорий (квот), которые должны быть представлены в заданных пропорциях. Например, при опросе покупателей универмага может быть запланировано провести отбор 150 респондентов, в том числе 90 женщин, из низ 25 — девушек, 20 — молодых женщин с маленькими детьми, 35 — женщин среднего возраста, одетых в деловой костюм, 10 — женщин старшего возраста; кроме того, планировался опрос 60 мужчин, из них 25 — подростков и юношей, 10 — молодых мужчин с детьми, 15 — мужчин, которые одеты в костюмы, 10 — мужчин, одетых в спортивную одежду. Для определения потребительских ориентаций и предпочтений такая выборка, может быть, и хороша, но если мы захотим по ней установить среднюю сумму покупок, их структуру, получим непредставительные результаты. Это происходит потому, что квотная выборка нацелена на отбор определенных категорий.

Выборка может быть нерепрезентативной, даже если она формируется в соответствии с известными пропорциями генеральной совокупности, но отбор проводится без какой-либо схемы — единицы набираются, как угодно, лишь бы обеспечить соотношение их категорий в тех же пропорциях, что и в генеральной совокупности (например, соотношение мужчин и женщин, респондентов в возрасте моложе и старше трудоспособного, в трудоспособном и т.д.).

Эти замечания должны предостеречь вас от подобных подходов к формированию выборки и еще раз показать необходимость объективного отбора.

7.3. Ошибка выборки

Все ошибки выборочного наблюдения подразделяются на ошибки выборки (случайные); ошибки, вызванные отклонением от схемы отбора (неслучайные); ошибки наблюдения (случайные и неслучайные). Плохо, когда ошибка выборки превышает допустимый размер погрешности, но слишком высокая точность также подозрительна и, как правило, свидетельствует об ошибках отбора.

К неслучайным ошибкам приводят ошибки отбора. Так бывает, если объективный отбор подменяется «удобной» выборкой. Например, когда появляются добровольные респонденты — те, кто сами предлагают, чтобы их опросили. Очевидно, что характеристики таких добровольцев и недобровольцев могут быть различны и это приведет к ошибочному заключению о генеральной совокупности.

Такая же опасность возникает при замене по какой-либо причине единиц, попавших в выборку, другими единицами (например, вместо отобранного домохозяйства, где в момент прихода интервьюера никто не открыл дверь, был проведен опрос в соседней квартире или интервьюер встретил решительный отказ участвовать в опросе и был вынужден пойти на замену домохозяйства). Как отмечает социолог В. И. Паниотто, систематические ошибки представляют собой некоторое постоянное смещение, которое не уменьшается с увеличением числа опрошенных и вызваны недостатками и просчетами в системе отбора респондентов. Если, например, для изучения общественного мнения жителей города в архитектурном управлении получить сведения о жилом фонде и из всех имеющихся в городе квартир отобрать случайным образом 400, а затем предложить интервьюерам опросить всех, кого они застанут в момент посещения в этих квартирах, то полученные данные не будут репрезентативны. Допущена систематическая ошибка: более подвижная часть населения попадает в выборку в меньшей пропорции, а менее подвижная — в большей пропорции, чем в генеральной совокупности. Пен-

сионеров, например, можно чаще застать дома, чем студентов-вечерников. При увеличении выборки эта ошибка не устраняется: если мы проведем опрос в 800 квартирах или даже во всех квартирах города (сплошной опрос), то полученные данные будут репрезентативны для населения, находящегося дома в момент прихода интервьюера, а не для всех жителей города.

Неслучайные ошибки могут возникнуть из-за методов сбора данных: наличия вопросов, слишком болезненных для опрашиваемых (об отношении к властям, если опрашиваются беженцы или пострадавшие от стихийных бедствий, и т.д.), или неудачной формы задания вопроса (очень трудно сформулировать так, чтобы всем было все понятно), или времени опроса (например, на вопрос молодым родителям, не жалеют ли они о том, что у них есть дети, можно получить разное распределение ответов в зависимости от того, проводился ли опрос долгим зимним вечером, когда все утомлены приготовлением уроков, простудами и т.д., или прекрасным летним днем, когда дети находятся на даче, в оздоровительном лагере).

Случайные ошибки — те, которые изменяются по вероятностным законам. К случайным относится ошибка выборки.

Ошибка выборки или, иначе говоря, *ошибка репрезентативности* — это разница между значением показателя, полученного по выборке, и генеральным параметром. Так, ошибка репрезентативности выборочной средней равна: $\varepsilon_{\bar{x}} = \bar{x} - \mu$, выборочной относительной величины $\varepsilon_p = p - \pi$, дисперсии $\varepsilon_{s^2} = s^2 - \sigma^2$, коэффициента корреляции $\varepsilon_r = r - \rho$.

Если представить, что было проведено бесконечное число выборок равного объема из одной и той же генеральной совокупности, то показатели отдельных выборок образовали бы ряд возможных значений: выборочных средних величин $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, ...; относительных величин p_1, p_2, p_3, ...; дисперсий s_1^2, s_2^2, s_3^2, ..., и т.д. Каждая выборка имеет свою ошибку репрезентативности. Следовательно, можно построить ряды распределения выборок по величине ошибки репрезентативности для каждого показателя: для средней, относительной величины и т.д. В таких распределениях прослеживается тенденция к концентрации ошибок около центрального значения. Число выборок с той или иной величиной ошибки репрезентативности может быть симметрично или асимметрично относительно этого центрального значения. При бесконечно большом числе выборок получится кривая частот, которая представляет *кривую выборочного распределения*. Свойства таких распределений используются для получения статистических заключений, установления вероятности появления той или иной величины ошибки репрезентативности.

Рассмотрим *выборочное распределение средней величины*. Такое распределение будет являться нормальным или приближаться к нему по мере увеличения объема выборки независимо от того, имеет или не имеет нормальное распределение та генеральная совокупность, из которой взяты выборки. С увеличением числа выборок средняя для всех выборок будет приближаться к генеральной средней $\bar{\varepsilon}_{\bar{x}} = 0$. По выборочному распределению может быть рассчитана средняя квадратическая ошибка репрезентативности:

$$s_\varepsilon = \sqrt{\frac{\sum \varepsilon_i^2 f_i}{\sum f_i}}, \qquad (7.1)$$

где ε_i^2 — квадрат ошибки репрезентативности для i-й выборки;
f_i — число выборок с одинаковым значением выборочной средней.

Среднее квадратическое отклонение выборочных средних от генеральной средней называется *средней ошибкой выборочной средней*:

$$s_{\bar{x}} = \sqrt{\frac{\sum (\bar{x}_i - \mu)^2 f_i}{\sum f_i}}. \qquad (7.2)$$

Поскольку, как правило, генеральная средняя μ неизвестна, этой формулой нельзя воспользоваться. Кроме того, в социально-экономических исследованиях выборки из одной и той же совокупности не проводятся многократно. Используют следующее соотношение: квадрат средней ошибки (дисперсия выборочных средних) прямо пропорционален дисперсии признака x в генеральной совокупности σ^2 и обратно пропорционален объему выборки n:

$$s_{\bar{x}}^2 = \frac{\sigma^2}{n}. \qquad (7.3)$$

Соответственно средняя ошибка выборочной средней равна:

$$s_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}. \quad (7.4)$$

Следовательно, средняя ошибка выборки тем больше, чем больше вариация в генеральной совокупности, и тем меньше, чем больше объем выборки.

Таким образом, можно утверждать, что отклонение выборочной средней $\bar{x}$ от генеральной средней μ в среднем равно $\pm s_{\bar{x}}$. Ошибка конкретной выборки $\Delta_{\bar{x}}$ может принимать различные значения, но отношение ее к средней ошибке практически не превышает ± 3, если величина n достаточно большая ($n > 100$). Отношение ошибки конкретной выборки к средней квадратической ошибке называется *нормированным отклонением* и обозначается как t:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}. \quad (7.5)$$

Распределение нормированного отклонения выборочной средней от генеральной средней при численности выборки $n \to \infty$ определяется уравнением Лапласа—Гаусса:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \quad (7.6)$$

где $f(t)$ — плотность вероятности;

σ — среднее квадратическое отклонение значений переменной x от средней в генеральной совокупности;

π и e — математические константы ($\pi = 3{,}14$, $e = 2{,}718$).

Поскольку средняя нормированных отклонений $t = 0$, дисперсия $\sigma_t^2 = 1$, т.е. $\sigma = 1$, выражение (7.6) может быть записано как

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \quad (7.7)$$

Уравнение (7.7) называют *стандартным уравнением нормальной кривой*. Величина $f(t)$ достигает максимума при $t = 0$, в этом случае $e^{t^2/2} = 1$. По мере увеличения t величина $e^{t^2/2}$

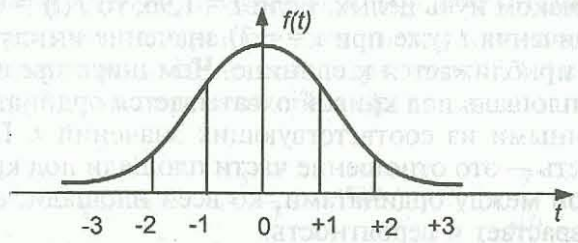


Рис. 7.1. Распределение ошибок выборочных средних

уменьшается и соответственно уменьшается $f(t)$. На рис. 7.1 приведен график кривой нормального распределения стандартизованных ошибок выборочных средних, t. Ординаты на графике соответствуют плотностям вероятностей при том или ином значении t. Для того чтобы определить вероятность значений в интервале от t_1 до t_2, следует найти отношение части площади кривой, заключенной между ординатами, соответствующими t_1 и t_2, ко всей площади кривой. Вся площадь под кривой нормального распределения вероятностей принимается за единицу.

Уравнение Лапласа—Гаусса предполагает непрерывное изменение t и неограниченное возрастание n. Поэтому площадь нормальной кривой, заключенную между ординатами t_1 и t_2, определяют, интегрируя функцию (7.7).

Имеются таблицы, которые содержат значения вероятностей для нормированных отклонений t или для интервалов от t_1 до t_2. Одна из таких таблиц приведена в приложении (табл. П.1). Эта таблица содержит пропорциональные доли площадей, заключенных между ординатами, соответствующими $\pm t$. Зная нормированное отклонение t, можно определить вероятность или на основе определенной вероятности установить величину t.

На пересечении строк и граф таблицы интеграла вероятностей находится значение вероятности $F(t)$, соответствующее данному значению t. Для краткости записи в таблице приводятся только десятичные знаки вероятности, следовательно, к табличному значению $F(t)$ надо приписывать нуль целых. Например, чтобы определить, какая вероятность соответствует $t = 1{,}96$, надо взять строку 1,9 и графу 6 и на их пересечении прочитать значение вероятности, добавив перед

первым знаком нуль целых. Если $t = 1{,}96$, то $F(t) = 0{,}9500$. По мере увеличения t (уже при $t = \pm 3$) значение интеграла вероятностей приближается к единице. Чем шире пределы t, тем большая площадь под кривой охватывается ординатами, восстановленными из соответствующих значений t. Поскольку вероятность — это отношение части площади под кривой, заключенной между ординатами, ко всей площади, соответственно возрастает и вероятность.

Распределение ошибок выборочных средних подчиняется закону нормального распределения или приближается к нему даже в случаях, когда генеральная совокупность имеет иную форму распределения.

Из формулы (7.5) следует, что отклонение выборочной средней от генеральной средней равно:

$$\Delta = \bar{x} - \mu = t\, s_{\bar{x}}. \qquad (7.8)$$

Нормированное отклонение t может быть установлено по табл. П.1 приложения. Для этого необходимо принять определенный уровень вероятности суждения о точности данной выборки.

Вероятность, которая принимается при расчете ошибки выборочной характеристики, называют *доверительной*. Чаще всего принимают доверительную вероятность равной 0,95, 0,954, 0,997 или даже 0,999. Уровень доверительной вероятности 0,95 означает, что только в 5 случаях из 100 ошибка может выйти за установленные границы; при вероятности 0,954 — в 46 случаях из 1000, при 0,997 — в 3 случаях, а при 0,999 — в 1 случае из 1000.

Для того чтобы вычислить ошибку выборки при принятой доверительной вероятности, нужно рассчитать величину средней ошибки выборки $s_{\bar{x}}$. Формула для ее определения (7.4) включает дисперсию признака в генеральной совокупности σ^2, которая, как правило, неизвестна. Может быть определена только выборочная дисперсия s^2. Доказано, что соотношение между σ^2 и s^2 определяется следующим равенством:

$$s^2 = \frac{n-1}{n} \cdot \sigma^2. \qquad (7.9)$$

Отсюда:

$$\sigma^2 = s^2 \frac{n}{n-1}. \qquad (7.10)$$

Если n велико, то сомножитель $n/(n-1) \approx 1$ и можно принять выборочную дисперсию в качестве оценки величины генеральной дисперсии. Подставив выражение (7.10) в формулу средней ошибки выборочной средней, получим:

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n-1}} = \frac{s}{\sqrt{n-1}} \quad \text{или} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}. \qquad (7.11)$$

Соответственно

$$\Delta_{\bar{x}} = t s_{\bar{x}} = t \frac{s}{\sqrt{n}}. \qquad (7.12)$$

Пример. Для определения скорости расчетов с кредиторами предприятий одного треста была проведена случайная выборка 50 платежных документов, по которым средний срок перечисления денег оказался равен 28,2 дня со стандартным отклонением 5,4 дня. Определим средний срок прохождения всех платежей в течение данного года с доверительной вероятностью $F(t) = 0{,}95$. Тогда $t = 1{,}96$; скорректированная дисперсия

$$s^2 = 5{,}4^2 \frac{50}{50-1} = 29{,}755;$$

средняя ошибка выборки

$$s_{\bar{x}} = \sqrt{\frac{29{,}755}{50}} = \sqrt{0{,}5951} = \pm 0{,}77 \text{ дня}.$$

Отклонение выборочной средней от генеральной с вероятностью 0,95 составит: $\Delta_x = 1{,}96 \cdot 0{,}77 = \pm 1{,}51$ дня.

Величина Δ называется доверительной ошибкой выборки или *предельной ошибкой выборки*. Рассчитав величину Δ, мы можем записать следующее неравенство:

$$28{,}2 - 1{,}51 \leqslant \mu \leqslant 28{,}2 + 1{,}51;$$

$$26{,}7 \text{ дня} \leqslant \mu \leqslant 29{,}7 \text{ дня}.$$

Таким образом, с вероятностью 0,95 можно утверждать, что средняя продолжительность расчетов предприятия данного треста с кредиторами составляет не менее 26,7 дня и не более 29,7 дня.

Ошибка выборки для *выборочной относительной величины (доли)* определяется аналогично. Дисперсию относительной величины определим по данным выборки:

$$s^2 = p(1-p), \qquad (7.13)$$

где p — доля тех или иных единиц в выборке.

Выражение (7.13) получено в соответствии с обычной формулой дисперсии. Поскольку имеется в виду альтернативная, или дихотомическая, переменная, обозначим ее значение в одной категории единиц 0, в другой —1. Тогда среднее значение переменной составит:

$$\frac{f_1 \cdot 0 + f_2 \cdot 1}{f_1 + f_2} = \frac{f_2}{n} = p;$$

квадрат отклонения от средней

$$s^2 = \frac{(0-p)^2 f_1 + (1-p)^2 f_2}{f_1 + f_2} = \frac{p^2 f_1 + f_2 - 2pf_2 + p^2 f_2}{n} =$$

$$= \frac{(f_1 + f_2) + f_2(1-2p)}{n} = p^2 + p(1-2p) = p^2 + p - 2p^2 =$$

$$= p(1-p),$$

что соответствует выражению (7.13).

Средняя ошибка выборочной доли

$$s_p = \sqrt{\frac{p(1-p)}{n}}. \qquad (7.14)$$

Предельная ошибка выборочной доли с принятой доверительной вероятностью имеет вид:

$$\Delta_p = t \cdot s_p = t\sqrt{\frac{p(1-p)}{n}}. \qquad (7.15)$$

Пример. По данным выборочного изучения 100 платежных документов предприятий одного треста оказалось, что в шести случаях сроки расчетов с кредиторами были превышены. С вероятностью 0,954 требуется установить доверительный интервал доли платежных документов треста без нарушения сроков:

$$q = \frac{6}{100} = 0{,}06, \text{ или } 6\%, \; p = 0{,}94;$$

$$s_p = \sqrt{\frac{0{,}94 \cdot 0{,}06}{100}} = \pm 0{,}024;$$

$$\Delta_p = 2 \cdot 0{,}024 = \pm 0{,}048.$$

Генеральная доля платежных документов π, не выходящих за установленные сроки, с вероятностью 0,954 находится в интервале

$$0{,}892 \leqslant \pi \leqslant 0{,}988, \text{ или } 89{,}2\% \leqslant \pi \leqslant 98{,}8\%.$$

7.4. Влияние вида выборки на величину ошибки выборки

Как указывалось в подразд. 7.2, при проведении выборочного наблюдения используются различные способы формирования выборочной совокупности: случайный отбор — повторный или бесповторный, механический, серийный, типический. Вид выборки влияет на величину ошибки выборки. При бесповторном отборе формулы средней ошибки выборки (7.4) и (7.14) дополняются множителем

$$\sqrt{\frac{N-n}{N-1}},$$

который корректирует величину ошибки выборки на объем генеральной совокупности и вероятность попадания единиц в выборку.

В серийной выборке дисперсия определяется как колеблемость между сериями:

$$s_x^2 = \frac{\sum_{r} (\breve{x}_j - \bar{x})^2}{r}, \qquad (7.16)$$

где r — число отобранных серий;

$\breve{x}_j$ — среднее значение признака x в j-й серии;

$\bar{x}$ — среднее значение в целом по выборке.

Формула (7.16) предполагает равенство серий по числу единиц, если это условие не выполняется, то в числитель выражения (7.16) вводится вес — число единиц в j-й серии, f_j; тогда в знаменателе указывается не r, а Σf_j. Межсерийная дисперсия представляет часть общей дисперсии признака x, и потому ее использование направлено на уменьшение ошибки выборки. Однако значение r намного меньше n, так как число отобранных гнезд намного меньше числа единиц наблюдения. Этот фактор увеличивает ошибку выборки. Его действие более значительно, нежели понижающее влияние межсерийной дисперсии — в результате ошибка серийной выборки в среднем больше ошибки выборки при отборе единицами.

При типическом отборе (стратифицированная, или районированная, выборка) дисперсия рассчитывается как средняя из внутрирайонных дисперсий:

$$\overline{s^2} = \frac{\sum_{1}^{m} s_{x_j}^2 n_j}{\sum_{1}^{m} n_j}, \qquad (7.17)$$

где m — число районов;

$s_{x_j}^2$ — выборочная дисперсия признака x в j-м районе.

$$s_{x_j}^2 = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 / n_j;$$

n_j — объем выборки в j-м районе.

Очевидно, что по правилу сложения дисперсий величина $\overline{s^2}$ меньше, чем величина общей дисперсии, s^2 на величину межрайонной дисперсии.

Величина ошибки районированной выборки меньше величины ошибки простой (нерайонированной выборки).

Часто используется сочетание районированного отбора с отбором сериями. Такой вид выборки обеспечивает преимущества в организации выборки и уменьшение ошибки выборки. Дисперсия такой выборки представляет среднюю из межсерийных дисперсий для каждого j-го района:

$$\overline{s_{\bar{x}}^2} = \frac{\sum_{j=1}^{m} s_{\bar{x}_j}^2 r_j}{\sum_{1}^{m} r_j}, \qquad (7.18)$$

где $s_{\bar{x}_j}^2$ — межсерийная дисперсия в j-м районе; $s_{\bar{x}_j}^2 = \sum_{1}^{r} (\bar{x}_{ij} - \bar{x}_j)^2 / r_j$;

m — число районов;

$\bar{x}_{ij}$ — средняя в i-й серии j-го района;

$\bar{x}_j$ — средняя в j-м районе;

r_j — число серий, отобранных в j-м районе.

Табл. 7.2 содержит формулы средней ошибки выборки для выборочной средней и выборочной относительной величины для разных видов выборки. В приведенных формулах требуют

Таблица 7.2

Формулы средней ошибки выборочной средней и выборочной относительной величины

Вид выборки	Средняя ошибка	
	выборочной средней	выборочной относительной величины (доли)
Повторная — отбор единицами	$\sqrt{\frac{s^2}{n}}$	$\sqrt{\frac{p(1-p)}{n}}$
Бесповторная — отбор единицами	$\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N-1}\right)}$	$\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$
Серийная (нерайонированная)	$\sqrt{\frac{\overline{s_{\bar{x}}^2}}{r}\left(\frac{R-r}{R-1}\right)}$	$\sqrt{\frac{\overline{s_p^2}}{r}\left(\frac{R-r}{R-1}\right)}$
Районированная — отбор единицами, бесповторная	$\sqrt{\frac{\overline{s^2}}{n}\left(\frac{N-n}{N-1}\right)}$	$\sqrt{\frac{\overline{s_p^2}}{n}\left(\frac{N-n}{N-1}\right)}$
Районированная — отбор сериями, бесповторная	$\sqrt{\frac{\overline{s_{\bar{x}}^2}}{r}\left(\frac{R-r}{R-1}\right)}$	$\sqrt{\frac{\overline{s_p^2}}{r}\left(\frac{R-r}{R-1}\right)}$

пояснения выражения дисперсий выборочной относительной величины.

При нерайонированной серийной выборке

$$s_p^2 = \frac{\sum_{j=1}^{r} (p_j - p)^2}{r}, \qquad (7.19)$$

где p_j — доля единиц определенной категории в j-й серии;
p — доля единиц этой категории в выборке.

При районированной серийной выборке

$$\overline{s_p^2} = \frac{\sum_{j=1}^{m} s_{p_j}^2 r_j}{\sum_{j=1}^{m} r_j}, \qquad (7.20)$$

где m — число районов;
$s_{p_j}^2$ — межсерийная дисперсия доли в j-м районе;
r_j — число серий, отобранных в j-м районе.

Пример. Рассмотрим влияние вида выборки на величину ошибки выборки (табл. 7.3). Предприятия легкой промышленности примем за генеральную совокупность. Ее характеристики:

Численность	$N = 60$
Генеральные средние:	$\mu_1 = 2{,}38$ оборота
	$\mu_2 = 2{,}23$
Генеральные дисперсии:	$\sigma_1^2 = 2{,}24$
	$\sigma_2^2 = 4{,}38$
Средние квадратические отклонения:	$\sigma^1 = 1{,}49$ оборота
	$\sigma^2 = 2{,}09$

Оборачиваемость запасов рассчитывается делением продолжительности периода (полгода) на среднюю продолжительность одного периода оборота запасов. Очевидно, что чем скорее оборачиваются запасы, тем выше их отдача. Коэффициент покрытия рассчитывается как отношение суммы всех источников покрытия запасов к стоимости запасов. Если значение этого показателя меньше единицы, то текущее финан-

Таблица 7.3

Показатели 60 предприятий легкой промышленности крупного города за I полугодие 2001 г.

№ п/п	Форма собственности	Оборачиваемость запасов, x_1	Коэффициент покрытия, x_2	№ п/п	Форма собственности	Оборачиваемость запасов, x_1	Коэффициент покрытия, x_2
1	Государственная	5,65	0,22	31	Частная	1,23	1,18
2	»	2,86	0,35	32	»	0,82	1,59
3	»	1,61	1,06	33	»	2,63	0,74
4	»	3,99	1,01	34	»	1,83	1,52
5	»	2,17	8,88	35	»	2,26	2,43
6	»	1,52	1,06	36	»	2,33	3,28
7	»	0,40	0,99	37	»	2,35	1,13
8	»	2,18	1,07	38	»	1,68	0,89
9	»	1,36	4,62	39	»	2,00	1,67
10	»	3,69	1,40	40	»	2,64	1,48
11	»	0,45	1,24	41	»	2,75	1,51
12	Частная	1,00	1,16	42	»	3,29	5,96
13	»	2,05	2,00	43	»	1,26	1,38
14	»	2,36	1,43	44	»	1,90	2,39
15	»	4,90	1,76	45	»	3,27	3,62
16	»	3,12	1,26	46	»	3,49	0,46
17	»	1,36	1,89	47	»	2,92	1,26
18	»	1,56	12,36	48	Смешанная	3,22	0,78
19	»	4,84	1,23	49	»	2,61	1,67
20	»	1,23	3,26	50	»	5,17	0,95
21	»	0,81	2,22	51	»	8,63	0,96
22	»	0,70	1,16	52	»	1,06	2,51
23	»	0,87	1,21	53	»	2,13	3,49
24	»	0,20	1,45	54	»	2,03	1,22
25	»	1,71	4,04	55	»	1,82	2,92
26	»	1,83	2,07	56	»	3,12	1,54
27	»	1,32	0,69	57	»	0,77	0,97
28	»	1,95	1,97	58	»	4,15	0,93
29	»	1,46	1,31	59	»	3,62	1,34
30	»	2,96	5,32	60	»	3,89	3,51

совое состояние предприятия рассматривается как неустойчивое. В нашем примере вариация этого признака примерно в два раза превосходит вариацию предприятий по уровню оборачиваемости запасов: $v_2 = 147\%$, $v_1 = 62\%$.

Проведем 30%-ную выборку. Объем выборки составит: $n = 20$ предприятий. При формировании выборки методом механического отбора каждое третье предприятие попадет в выборку. Отбор начинаем с полушага отбора, т.е. первым предприятием, попавшим в выборку, является второе по списку. Средние по выборке равны: оборачиваемость запасов $\bar{x}_1 = 2{,}16$ оборота, коэффициент покрытия $\bar{x}_2 = 2{,}01$.

Средняя ошибка выборочной средней оборачиваемости запасов

$$s_{\bar{x}_1} = \sqrt{\frac{2{,}24}{20}\left(\frac{60-20}{60-1}\right)} = \pm 0{,}275 \text{ оборота.}$$

Средняя ошибка выборочного среднего коэффициента покрытия

$$s_{\bar{x}_2} = \sqrt{\frac{4{,}38}{20}\left(\frac{60-20}{60-1}\right)} = \pm 0{,}385.$$

С вероятностью 0,954 можно утверждать, что средняя оборачиваемость запасов на предприятиях легкой промышленности не ниже $\bar{x}_1 - 2s_{\bar{x}_1} = 2{,}16 - 0{,}55 = 1{,}61$ оборота и не выше $\bar{x}_1 + 2s_{\bar{x}_1} = 2{,}16 + 0{,}55 = 2{,}71$ оборота.

Действительно, генеральная средняя ($\mu_1 = 2{,}38$) попадает в этот интервал.

Фактическая ошибка репрезентативности

$$\varepsilon_1 = |\bar{x}_1 - \mu_1| = |2{,}16 - 2{,}38| = 0{,}22 \text{ оборота.}$$

Эта величина меньше предельной ошибки выборки, гарантированной с принятой доверительной вероятностью, — $0{,}22 < 0{,}55$. Следовательно, выборка репрезентативна по этому признаку.

Вычислим предельную ошибку выборки коэффициента покрытия и определим доверительный интервал для этой характеристики. Его нижняя граница с той же вероятностью:

$$\bar{x}_2 - 2s_{\bar{x}_2} = 2{,}01 - 0{,}77 = 1{,}24;$$

верхняя граница:

$$\bar{x}_2 + 2s_{\bar{x}_2} = 2{,}01 + 0{,}77 = 2{,}78.$$

Генеральная средняя ($\mu_2 = 2{,}23$) также попадает в доверительный интервал.

Фактическая ошибка репрезентативности составляет:

$$\varepsilon_2 = |\bar{x}_2 - \mu| = |2{,}01 - 2{,}23| = 0{,}22.$$

Эта величина меньше предельной ошибки выборки (0,77), что дает основание считать выборку репрезентативной и по этому признаку.

В генеральной совокупности доля единиц с неустойчивым финансовым положением ($x_2 < 1$) составила: $\pi = \frac{12}{60} = 0{,}20$, в выборке: $p = \frac{3}{20} = 0{,}15$.

Доверительный интервал для оценки доли таких предприятий в генеральной совокупности составляет с вероятностью 0,954:

$$0{,}15 \pm 2\sqrt{\frac{0{,}15 \cdot 0{,}85}{20} \cdot \left(\frac{60-20}{60-1}\right)};$$

$$0{,}15 \pm 0{,}13,$$

т.е. таких предприятий должно быть не меньше 2% и не больше 28%. Фактически в генеральной совокупности их оказалось 20% общего числа предприятий, т.е. выборка дает репрезентативный результат и по этому показателю.

Выполненная выборка формировалась как простая бесповторная механическая. Однако наверняка статистик будет стремиться учесть структуру генеральной совокупности, поэтому более естественной была бы выборка, учитывающая выделение предприятий разных форм собственности. Тогда выборка должна быть районированной.

Пример. Генеральная совокупность состоит из 11 государственных предприятий, 36 частных, 13 смешанных. В выборке эти пропорции соблюдаются следующим образом: отобраны по 4 предприятия государственных и смешанных и 12 — частных (табл. 7.4).

Таблица 7.4

Генеральные выборочные характеристики

Предприятия	Генеральные характеристики		Выборочные характеристики	
	средние	доли	средние	доли
Государственные	$\mu_1 = 2{,}35$	$\pi_1 = 0{,}27$	$\bar{x}_1 = 1{,}92$	$p_1 = 0{,}25$
Частные	$\mu_1 = 2{,}08$	$\pi_2 = 0{,}11$	$\bar{x}_1 = 2{,}26$	$p_2 = 0{,}08$
Смешанные	$\mu_1 = 3{,}25$	$\pi_3 = 0{,}31$	$\bar{x}_1 = 2{,}41$	$p_3 = 0{,}25$

Средняя из внутрирайонных дисперсий, рассчитанных по каждой группе предприятий в генеральной совокупности, составит

$$\overline{\sigma^2} = \frac{\sum \sigma_j^2 \cdot N_j}{\sum N_j} = \frac{25{,}32 + 39{,}16 + 50{,}7}{60} = 1{,}9197$$

(по данным выборки межгрупповая дисперсия равна $\bar{s}^2 = 1{,}063$).

Эта величина меньше общей дисперсии без учета районирования ($\sigma^2 = 2{,}24$). Следовательно, и величина ошибки выборки при районированном отборе будет меньше:

$$\Delta_{\bar{x}} = 2\sqrt{\frac{1{,}9197}{20} \cdot \left(\frac{60-20}{60-1}\right)} = 2 \cdot (\pm 0{,}255) = \pm 0{,}510.$$

Итак, с вероятностью 0,954 генеральная средняя оборачиваемости запасов находится в интервале $2{,}38 \pm 0{,}51$;

$$1{,}87 \leqslant \mu \leqslant 2{,}89.$$

Для того чтобы понять, насколько целесообразно применение районированного отбора в том или ином случае, можно воспользоваться корреляционным отношением η. Согласно правилу сложения дисперсий средняя из внутригрупповых дисперсий может быть представлена как

$$\bar{s}^2 = s^2(1 - \eta^2), \qquad (7.21)$$

где η^2 — квадрат корреляционного отношения, равный $\delta^2 : s^2$.

Следовательно, применение районированной (типической) выборки изменяет предельную ошибку на $\sqrt{1-\eta^2}$. В нашем примере для первой переменной (оборачиваемости) имеем:

$$\sqrt{1-\eta^2} = \sqrt{1 - \frac{1{,}9197}{2{,}2378}} = 0{,}37.$$

Сопоставим полученный результат с изменением предельной ошибки выборки: $\Delta_{\bar{x}}$ (без учета районирования) = 0,55; $\Delta_{\bar{x}}$ (при районировании) = 0,51, т.е. ошибка уменьшилась примерно на 7%.

Корреляционное отношение используется и при корректировке величины

$$t^* = t\sqrt{1-\eta^2}. \qquad (7.22)$$

Тогда при вероятности 0,954 и $t = 2$: $t^* = 2 \cdot \sqrt{0{,}86} = 1{,}85$, т.е. вместо $t = 2$ достаточно взять $t = 1{,}85$.

Многие выборки формируются как многоступенчатые. *Ошибка многоступенчатой выборки* может быть представлена как

$$\bar{x}_k - \mu = (\bar{x}_k - \bar{x}_{k-1}) + \ldots + (\bar{x}_2 - \bar{x}_1) + (\bar{x}_1 - \mu).$$

Она складывается из ошибок отдельных ступеней. Поэтому практически используется не больше 2—3 ступеней отбора.

Средняя ошибка выборки при двухступенчатом отборе рассчитывается по формуле

$$s_{\bar{x}} = \sqrt{\frac{s_{x_1}^2}{m} + \sum_1^m \frac{s_{x_2}^2}{n_i}\frac{1}{m}}, \qquad (7.23)$$

где m — число отобранных «крупных» единиц;

$s_{x_1}^2$ — дисперсия признака x по совокупности «крупных» единиц;

$s_{x_2}^2$ — дисперсия признака x в каждой из отобранных «крупных» единиц;

n_i — число отобранных единиц наблюдения в i-й «крупной» единице.

Таким образом, использование многоступенчатой выборки улучшает организацию выборки, но увеличивает ее ошибку.

Кроме рассмотренных применяется *многофазовая выборка*, когда одни сведения собираются по всем единицам выборки, а другие — только по подвыборке из первоначальной выборки.

При периодическом повторении выборочных обследований с целью изучения динамики явлений применяются либо *независимые выборки* — через определенные промежутки времени отбор каждый раз проводится независимо от предыдущих выборок; либо *фиксированные выборки* — в этом случае повторные обследования проводятся по одной и той же выборке. В связи с тем, что в фиксированной выборке могут происходить изменения (прежде всего за счет выбытия единиц), практикуют периодическую адаптацию фиксированной выборки к происходящим изменениям. Чаще для целей изучения динамики используется промежуточный вариант — *ротационная выборка* (частичное замещение). При этом нужно следовать определенному плану замещения, например, каждый раз замещать четверть выборки, тогда каждая первоначально отобранная единица останется в четырех следующих друг за другом выборках.

Названные виды выборок ориентированы на отбор конкретных материальных явлений. Помимо них следует назвать как особый вид выборки *метод моментных наблюдений*. Сущность этого метода состоит в периодической фиксации состояний наблюдаемых единиц в отобранные моменты времени. Расчет объема такой выборки дает количество моментов. Этот вид выборочного наблюдения применяется при изучении использования производственного оборудования либо рабочего времени (подразд. 7.7).

7.5. Задачи, решаемые при применении выборочного метода

При использовании выборочного метода возникают три основные задачи:

- определение объема выборки, необходимого для получения требуемой точности результатов с заданной вероятностью;
- определение возможного предела ошибки репрезентативности, гарантированного с заданной вероятностью, и сравнение его с величиной допустимой погрешности;
- определение вероятности того, что ошибка выборки не превысит допустимой погрешности.

Все эти задачи решаются на основе теоремы Чебышева, согласно которой $P\{|\bar{x} - \mu| < \varepsilon\} \geqslant 1 - h$, когда n — достаточно большое число; ε и h — сколь угодно малые положительные числа. Это соотношение, как было показано в подразд. 7.3, может быть выражено через формулу предельной ошибки выборки: $\Delta_x = ts_x$ или $\Delta_p = ts_p$. Решение указанных задач зависит от того, какие величины в формуле предельной ошибки заданы, а какие нужно найти.

Объем выборки рассчитывается на стадии проектирования выборочного обследования. Так как

$$\Delta = t\sqrt{\frac{\sigma^2}{n}},$$

то

$$n = \frac{t^2\sigma^2}{\Delta^2}, \qquad (7.24)$$

где Δ — допустимая погрешность, которая задается исследователем исходя из требуемой точности результатов проектируемой выборки;

t — табличная величина, соответствующая заданной доверительной вероятности $F(t)$, с которой будут гарантированы оценки генеральной совокупности по данным выборочного обследования;

σ^2 — генеральная дисперсия.

Последняя величина, как правило, неизвестна. Используются какие-либо ее оценки: результаты прошлых обследований той же совокупности, если ее структура и условия развития достаточно стабильны, или же зная примерную величину средней, находят дисперсию из соотношения

$$\sigma \approx \frac{1}{3}\bar{x},$$

если известны x_{max} и x_{min}, то можно определить среднее квадратическое отклонение в соответствии с правилом «трех сигм»:

$$\sigma = \frac{1}{6}(x_{max} - x_{min}),$$

так как в нормальном распределении в размахе вариации «укладывается» 6σ ($\pm 3\sigma$). Если распределение заведомо асимметричное, то

$$\sigma \approx \frac{1}{5}(x_{max} - x_{min}).$$

Для относительной величины принимают максимальную величину дисперсии: $\sigma^2_{max} = 0{,}5 \cdot 0{,}5 = 0{,}25$.

При расчете n не следует гнаться за большими значениями t и малыми значениями Δ, поскольку это приведет к увеличению объема выборки и, следовательно, к увеличению затрат средств, труда и времени, вовсе не являющемуся необходимым.

Формула (7.24) не учитывает бесповторности отбора и дает максимальную величину выборки, которую можно скорректировать «на бесповторность». Так как

$$\Delta = t\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},$$

то на основе формулы (7.24) получаем выражение скорректированного объема выборки (n):

$$n = \frac{n_0}{\frac{n_0 + (N-1)}{N}}, \qquad (7.25)$$

где $n_0 = \frac{t^2\sigma^2}{\Delta^2}$.

При больших размерах генеральной совокупности скорректированный объем выборки незначительно отличается от n_0.

Например, для изучения структуры и стоимости покупок в универмаге из 10 000 покупателей следует отобрать то число человек, которое бы обеспечивало с вероятностью 0,95 определение средней стоимости покупок с точностью не менее 50 руб. Дисперсию примем равной 62 500 (по прошлому обследованию):

$$n_0 = \frac{1{,}96^2 \cdot 62\,500}{2500} = 96{,}04 \approx 97 \text{ человек};$$

тогда скорректированная численность

$$n = \frac{97}{\frac{97 + (10000 + 1)}{10000}} = 96 \text{ человек } (\approx 100 \text{ человек}),$$

т.е. в данном случае корректировка не оказала влияния на результаты расчета. Все округления даются с превышением. Окончательный результат: должно быть опрошено 100 человек.

При проектировании районированной выборки рассчитанный объем выборки распределяют пропорционально численности районов (*пропорциональный отбор*):

$$n_i = n \cdot \frac{N_i}{N}, \qquad (7.26)$$

где n_i — объем выборки для i-го района;
n — общий объем выборки;
N_i — объем i-го района в генеральной совокупности;
N — общий объем генеральной совокупности.

При различиях в однородности выделенных районов лучшие результаты дает распределение запланированного объема выборки между районами не только с учетом их объема, но и с учетом дисперсии признака (*оптимальный отбор*). В этом случае объем выборки в i-м районе определяется как

$$n_i = n \cdot \frac{N_i\sigma_i^2}{\sum_{i=1}^{m} N_i\sigma_i^2}, \qquad (7.27)$$

где σ_i^2 — дисперсия признака x в i-м районе.

При любом виде проектируемой выборки расчет объема выборки начинают по формуле повторного отбора (7.24). Если в результате расчета n доля отбора превысит 5%, рассчитывают объем выборки по формуле бесповторного отбора либо по формуле (7.25), либо как

$$n = \frac{Nt^2\sigma^2}{N\Delta^2 + t^2\sigma^2}. \qquad (7.28)$$

Если доля отбора меньше 5%, к формуле бесповторного отбора не переходят, так как это несущественно скажется на величине n (как это было в рассмотренном примере).

Выборка должна быть такой, чтобы выборочные показатели по всем основным характеристикам были репрезентативны. Поэтому численность выборки рассчитывают многократно исходя из допустимых ошибок разных показателей, значения которых в генеральной совокупности известны.

Например, при выборочном учете детей школьного возраста требуется определить число семей, которые надо обследовать. При этом надо учесть: а) число детей в возрасте 6—7 лет; б) число детей в возрасте 6—15 лет; в) число детей в возрасте 16—17 лет; г) среднедушевой доход (например, для решения вопроса о строительстве базы отдыха).

Поскольку репрезентируемые признаки могут иметь разную размерность, допустимая погрешность для каждого из них задается в виде относительной величины ($\Delta : \bar{x}$) (например, планируется, что в определении среднего размера семьи ошибка должна быть не больше 2%, в определении дохода — не больше 3% и т.д.). В этом случае вместо дисперсии в формуле (7.24) берется квадрат коэффициента вариации, т.е. $n = t^2v^2/\Delta : x$.

Вычислив значение n, на основе каждой из характеристик получаем разные объемы выборки: 1200; 300; 700; 100. Обследовать необходимо 1200 семей, т.е. из рассчитанных численностей берется максимальная. При резких различиях необходимых объемов выборки для разных вопросов программы проводится многофазный отбор. В рассмотренном примере среднедушевой доход достаточно учитывать в одной из каждых 12 семей, попавших в выборку.

Многофазный отбор, как правило, довольно сложно организовать, может быть нарушен принцип случайности отбора. Поэтому для обеспечения репрезентативности оказывается выгоднее затратить больше средств на учет большего числа единиц совокупности. Многофазный отбор целесообразно применять, если соотношение между рассчитанными объемами выборки по крайней мере 1 : 6.

Поскольку расчет необходимой численности выборки основан не на точных, а на предположительных данных о колеблемости в совокупности, надо соблюдать следующие рекомендации: абсолютную величину n округлять только вверх; долю отбора округлять только вниз, т.е. из предосторожности планировать несколько бóльший объем выборки, чем показывают расчеты.

Объем многоступенчатой выборки рекомендуется увеличить не менее чем на 10% рассчитанной численности, поскольку, как было показано в подразд. 7.4, многоступенчатость отбора увеличивает ошибку выборки.

После проведения выборки рассчитывают ошибки выборочных показателей (ошибки репрезентативности), которые используются для оценки результатов выборки и для получения характеристик генеральной совокупности.

Пример. На электроламповом заводе взято для проверки 100 ламп. Средняя продолжительность их горения оказалась 1420 ч со средним квадратическим отклонением 61,03 ч. Поскольку приемщика продукции интересует качество всей партии (50 тыс. электроламп), оценивают точность полученной средней. Средняя возможная ошибка вычисленной выборочной средней:

$$s_{\bar{x}} = \frac{61{,}03}{\sqrt{100}} = \pm 6{,}1 \text{ ч.}$$

С вероятностью 0,954 предел возможной ошибки:

$$\Delta_{\bar{x}} = 2 \cdot 6{,}1 = \pm 12{,}2 \text{ ч.}$$

С вероятностью 0,954 можно утверждать, что средняя продолжительность горения одной электролампы во всей партии будет находиться в пределах от 1407,8 до 1432,2 ч; 46 электроламп из 1000 могут иметь срок горения, выходящий за эти пределы.

Приемщика продукции интересуют отклонения от вычисленных пределов только в сторону сокращения продолжительности горения. Меньше 1407,8 ч могут гореть 23 лампы из 1000. На основе этого приемщик продукции решает вопрос о годности всей партии электроламп.

Формулировка вопроса может быть уточнена: у какой доли ламп срок службы окажется меньше установленного лимита? Для потребителя таким лимитом являются 1410 ч, продукция с меньшим сроком горения неприемлема.

При контрольной проверке 100 ламп 10 ламп горели менее 1410 ч, их удельный вес (доля) $p = 0{,}1$, или 10%. Средняя возможная ошибка этой доли

$$s_p = \sqrt{\frac{0{,}1 \cdot 0{,}9}{100}} = \pm 0{,}03, \text{ или } \pm 3\%.$$

С вероятностью 0,954 предел ошибки доли $\Delta_p = 2 \cdot 0{,}03 = \pm 0{,}06$, или ±6%. Следовательно, во всей партии можно ожидать долю некачественных электроламп от 4 до 16%.

Чаще всего делают заключение об удовлетворительности выборки, сопоставляя полученные пределы ошибок выборочных показателей с величинами допустимых погрешностей. Может получиться, что предел ошибки, рассчитанный с заданной вероятностью, окажется выше допустимого размера погрешности. В этих случаях определяют вероятность того, что ошибка выборки не превзойдет допускаемую погрешность. Решение этой задачи и заключается в отыскании $F(t)$ на основе формулы предела ошибки выборки:

$$\Delta = t\sqrt{\frac{s^2}{n}},$$

где Δ — допустимый размер погрешности оцениваемого показателя;
s^2 — дисперсия показателя, рассчитанная по данным выборочного наблюдения;
n — объем проведенной выборки.

Вернемся к нашему примеру, в котором рассматривается оценка качества электроламп. Если при приемке партии электроламп ставится условие, что минимальный срок горения 1410 ч, то, учитывая среднюю продолжительность горения по выборке ($\bar{x} = 1420$ ч), допустимая погрешность равна 10 ч: $1410 - 1420 = -10$ ч.

Как было установлено выше, с вероятностью 0,954 предел возможной ошибки выборочной средней составил 12,2 ч, что превосходит допустимую погрешность. Является ли это основанием для браковки всей партии? Для ответа на этот вопрос определяют вероятность риска при приемке продукции:

$$10 = t\sqrt{\frac{61{,}03^2}{100}}, \text{ отсюда } t = 1{,}64.$$

Соответствующая доверительная вероятность 0,899 (приложение, табл. П.1). Вероятность того, что средний срок горения лампы меньше 1410 ч, равна:

$$1 - \frac{1 + F(t)}{2} = 1 - (0{,}5 + \frac{0{,}899}{2}) = 0{,}05.$$

Следовательно, из 100 ламп 5 могут гореть менее 1410 ч — риск появления некачественной продукции достаточно высок.

Аналогично можно определить вероятность того, что предел ошибки доли не превысит допустимую погрешность.

Оценки надежности выборочных показателей, как показано на примере, позволяют принять обоснованные решения в отношении генеральной совокупности.

7.6. Распространение данных выборочного наблюдения на генеральную совокупность

Конечной целью выборочного наблюдения является характеристика генеральной совокупности на основе данных, полученных по выборке. При этом исходят из того, что все средние и относительные показатели, полученные по выборке, являются несмещенными и эффективными характеристиками генеральной совокупности.

Выборочные средние и относительные величины распространяются на генеральную совокупность обязательно с учетом предела их возможной ошибки. Приводится выборочный показатель со справкой о пределах ошибки с указанием доверительной вероятности: $\bar{x} \pm \Delta_{\bar{x}}$, $p \pm \Delta_p$. Или же указывают границы значений генеральной характеристики с определенной вероятностью $F(t)$:

$$\bar{x} - \Delta_{\bar{x}} \leqslant \mu \leqslant \bar{x} + \Delta_{\bar{x}};$$
$$p - \Delta_p \leqslant \pi \leqslant p + \Delta_p.$$

Последняя форма записи является основной.

Иногда требуется указать только один (верхний или нижний) предел характеристики генеральной совокупности. При испытании качества продукции часто нас не интересуют положительные ошибки выборки (качество фактически выше, чем получилось по выборке), беспокоит нижний предел, как в примере, рассмотренном в подразд. 7.5. В некоторых случаях, напротив, интерес вызывают верхние границы оцениваемых показателей, например при анализе расхода материалов. Так что при характеристике генеральной совокупности всегда указывают неблагоприятный предел.

На основе выборки могут быть получены и значения объемных показателей, т.е. подсчетов для генеральной совокупности. Такой расчет осуществляется двумя способами: прямым расчетом и способом коэффициентов. *Прямой расчет* за-

ключается в том, что выборочная средняя или доля умножается на объем генеральной совокупности:

$$\bar{x}N = \sum_{1}^{N} x_i.$$

Поскольку средняя величина имеет ошибку репрезентативности $\pm\Delta_{\bar{x}}$, можно считать, что итоговый подсчет в генеральной совокупности находится в пределах

$$(\bar{x} - \Delta_{\bar{x}})N \leqslant \sum_{i=1}^{N} x_i \leqslant (\bar{x} + \Delta_{\bar{x}})N. \qquad (7.29)$$

Итоговый подсчет по генеральной совокупности можно получить на основе *итогового подсчета по выборке*, разделив его величину на долю отбора единиц совокупности

$$\sum_{1}^{n} x/(n/N).$$

Прежде чем производить расчет объемных показателей для генеральной совокупности, нужно убедиться, что структура выборки соответствует структуре генеральной совокупности. При наличии значительных смещений в структуре выборки в долях отдельных групп (0,03 и выше) следует применить *метод перевзвешивания*, т.е. рассчитывать генеральную среднюю на основе выборочных средних по группам и удельного веса этих групп в генеральной совокупности:

$$\mu = \bar{x}_1 w_1 + \ldots + \bar{x}_m w_m,$$

где $w_i = N_i/N$.

При *способе коэффициентов* также используются не только выборочные данные, но и сведения о генеральной совокупности.

Этот способ основан на связи признаков друг с другом. Например, в результате выборочного обследования семей города получены размер среднедушевого дохода ($\bar{x}$), средний доход семьи ($\bar{y}$) и среднее число человек в семье ($\bar{z}$). Так что $\bar{x} = \bar{y}/\bar{z}$.

Зная численность населения города N, требуется рассчитать общую величину денежного дохода населения. Очевидно, это можно сделать, умножив среднедушевой доход на общее число жителей в городе: $\bar{x}N$. Общий доход можно получить, суммируя доход отдельных семей; численность населения можно получить, суммируя данные о числе членов семей.

Тогда:

$$\sum_{1}^{N} x = \frac{\sum_{1}^{n} y}{\sum_{1}^{n} z} \sum_{1}^{N} z.$$

Средний душевой доход $\sum_{1}^{N} y / \sum_{1}^{N} z$ представляет собой коэффициент, который связывает две характеристики. Этот коэффициент рассчитывается как отношение двух итоговых подсчетов по выборке:

$$\sum_{1}^{n} y : \sum_{1}^{n} z.$$

Следовательно,

$$\sum_{1}^{N} x_i = \frac{\sum_{1}^{n} y_i}{\sum_{1}^{n} z_i} \cdot \sum_{1}^{N} z_i = \sum_{1}^{n} y_i \frac{\sum_{1}^{N} z_i}{\sum_{1}^{n} z_i}.$$

Последний сомножитель не что иное, как обратная величина доли отбора, рассчитанной по значениям признака z.

Значит, итоговый подсчет по генеральной совокупности может быть получен делением соответствующего итогового подсчета по выборке на долю отбора. При прямом расчете берется доля отбора единиц совокупности, при способе коэффициентов — доля отбора по значению какого-либо признака.

Эффективность способа коэффициентов по сравнению с методом прямого расчета зависит от того, насколько тесно связаны между собой признаки, лежащие в основе расчета коэффициента, т.е. признак, по которому подсчитывается итог, и признак, по которому определяется доля отбора. Эффект проявляется, если коэффициент корреляции между ними больше 0,8.

Способ коэффициентов используется для корректировки данных сплошного наблюдения. Например, по данным переписи скота была получена величина поголовья свиней в районе 10 000, в том числе в тех хозяйствах, которые потом были

охвачены контрольным обходом, сплошное наблюдение показало поголовье свиней 1100. Контрольный обход дал уточненную цифру: не 1100, а 1107 свиней. Тогда поправочный коэффициент

$$k = \frac{7}{1100} = 0{,}0064.$$

Отсюда скорректированная численность поголовья свиней во всем районе составляет:

$$N = N' + \Delta N;$$
$$\Delta N = kN' = 0{,}0064 \cdot 10000 = 64;$$
$$N = 10000 + 64 = 10\,064 \text{ головы.}$$

7.7. Малая выборка

Таблицы интеграла вероятностей используются для выборок большого объема из бесконечно большой генеральной совокупности. Но уже при $n < 100$ возникает несоответствие между табличными данными и вероятностью предела; при $n \leqslant 30$ погрешность становится значительной. Несоответствие обусловлено главным образом характером распределения единиц генеральной совокупности. При большом объеме выборки особенность распределения в генеральной совокупности не имеет значения, так как распределение отклонений выборочного показателя от генеральной характеристики при большой выборке всегда оказывается нормальным.

В выборках небольшого объема $n \leqslant 30$ характер распределения генеральной совокупности сказывается на распределении ошибок выборки. Поэтому для расчета ошибки выборки при небольшом объеме наблюдения (уже менее 100 единиц) отбор должен проводиться из совокупности, имеющей нормальное распределение.

Теория малых выборок разработана английским статистиком В. Госсетом (писавшим под псевдонимом Стьюдент) в начале XX в. В 1908 г. им построено специальное распределение, которое позволяет и при малых выборках соотносить t и доверительную вероятность $F(t)$. При $n > 100$ таблицы распределения Стьюдента дают те же результаты, что и таблицы интеграла вероятностей Лапласа, при $30 \leqslant n \leqslant 100$ различия незначительны. Поэтому практически к малым выборкам относят выборки объемом менее 30 единиц (безусловно большой считается выборка с объемом более 100 единиц).

Использование малых выборок в ряде случаев обусловлено характером обследуемой совокупности. Так, в селекционной работе «чистого» опыта легче добиться на небольшом числе делянок. Производственный или экономический эксперимент также легче провести на небольшом числе испытаний.

Как уже отмечалось, в случае малой выборки только для нормально распределенной генеральной совокупности могут быть рассчитаны и доверительные вероятности, и доверительные пределы генеральной средней.

Плотность вероятностей распределения Стьюдента описывается функцией

$$f(t, n) = B_n\left(1 + \frac{t^2}{n-1}\right)^{-n/2}, \qquad (7.30)$$

где t — текущая переменная;
n — объем выборки;
B_n — величина, зависящая лишь от n.

Распределение Стьюдента имеет только один параметр: $d.f.$ — число степеней свободы (англ. degrees of freedom), в отечественной литературе иногда обозначается буквой k.

Это распределение, как и нормальное, симметрично относительно точки $t = 0$, но оно более пологое. При увеличении объема выборки, а следовательно, и числа степеней свободы распределение Стьюдента быстро приближается к нормальному. Число степеней свободы равно числу тех индивидуальных значений признаков, которыми нужно располагать для определения искомой характеристики.

Так, для расчета дисперсии должна быть известна средняя величина. Поэтому при расчете дисперсии $d.f. = n - 1$.

Таблицы распределения Стьюдента публикуются в двух вариантах:

1) аналогично таблицам интеграла вероятностей приводятся значения t и соответствующие вероятности $F(t)$ при разном числе степеней свободы;

2) значения t приводятся для наиболее употребимых доверительных вероятностей 0,90; 0,95 и 0,99 или для $1 - 0{,}9 = 0{,}1$, $1 - 0{,}95 = 0{,}05$ и $1 - 0{,}99 = 0{,}01$ при разном числе степеней

свободы. Такого рода таблица приведена в табл. П.2 приложения, а также значение t-критерия Стьюдента при уровне значимости 0,10; 0,05; 0,01.

При малых выборках расчет средней возможной ошибки основан на выборочной дисперсии, поэтому

$$s_{\bar{x}} = \sqrt{\frac{s_x^2}{n-1}}. \qquad (7.31)$$

Приведенная формула используется для определения предела возможной ошибки выборочного показателя:

$$\Delta_{\bar{x}} = ts_{\bar{x}}.$$

Порядок расчетов тот же, что и при больших выборках.

Пример. Для изучения интенсивности труда было организовано наблюдение за 10 отобранными рабочими. Доля работавших все время была равной 0,40, дисперсия: 0,4 · 0,6 = = 0,24. По табл. П.2 приложения находим: $F(t) = 0{,}95$ и $d.f. = n - 1 = 9$, $t = 2{,}26$. Рассчитаем среднюю ошибку выборки доли работавших все время:

$$s_p = \sqrt{\frac{0{,}24}{10-1}} = \pm 0{,}16.$$

Тогда предельная ошибка выборки $\Delta_p = 2{,}26 \cdot 0{,}16 = \pm 0{,}36$. Таким образом, с вероятностью 0,95 доля рабочих, работавших без простоев, в данном цехе предприятия находится в пределах

$$39{,}64\% \leqslant \pi \leqslant 40{,}36\%$$

или

$$39{,}6\% \leqslant \pi \leqslant 40{,}4\%.$$

Если бы мы использовали для расчета доверительных границ генерального параметра таблицу интеграла вероятностей, то t было бы равно 1,96 и $\Delta_p = \pm 0{,}31$, т.е. доверительный интервал был бы несколько у́же.

Малые выборки широко применяются для решения задач, связанных с испытанием статистических гипотез, особенно гипотез о средних величинах.

7.8. Примеры применения выборочного метода

Потребность в использовании выборочного метода, выработке вероятностных суждений в современной отечественной практике непрерывно расширяется. В государственной статистике основными направлениями использования выборочного метода традиционно являются бюджетные обследования домохозяйств, выборочные переписи населения, контрольные обходы и проверки после проведения сплошных обследований.

Создание ЕГРПО, в котором фиксируются все хозяйствующие субъекты на территории Российской Федерации всех форм собственности, открывает возможность проведения разнообразных выборочных обследований в области экономики.

В области социальных исследований для государственной статистики главным является *бюджетное обследование*, которое охватывает примерно 45 тыс. домохозяйств. Оно основано на многоступенчатом отборе. Общий объем выборки распределяется по сферам занятости (для работающих) и территориям. Затем для работающих проводится отбор предприятий в пределах каждой отрасли в отобранной территории. Если, например, нужно отобрать 100 рабочих, занятых в определенной отрасли, для обследования семейных бюджетов так, чтобы на каждом отобранном предприятии было не менее 20 бюджетов, включающих рабочих с разным уровнем заработной платы, то, значит, должно быть отобрано: 100 : 20 = = 5 предприятий. Отбор предприятий проводят по списку, в котором предприятия располагаются в порядке убывания средней заработной платы рабочих, указываются общее число рабочих, их суммарная заработная плата. Шаг отбора определяется делением общего числа рабочих на предприятиях данной отрасли на число отбираемых предприятий. Если всего на предприятиях данной отрасли в области занято 30 525 человек, то шаг отбора равен: 30525 : 5 = 6105. По данным кумулятивной численности рабочих с рассчитанным шагом отбора проводится отбор предприятий, которые затем проверяются на репрезентативность по показателю средней месячной заработной платы. Следующая стадия связана с отбором рабочих на выбранных предприятиях: среди 20 бюджетов должны быть пропорционально представлены бюджеты семей малоквали-

фицированных и высококвалифицированных рабочих, а среди этих категорий отбор проводится механически по спискам рабочих, составленным в порядке убывания средней месячной заработной платы. Выборочная совокупность при бюджетных обследованиях включает и семьи неработающих (пенсионеров, студентов, инвалидов) и одиночек.

Задачей статистики в области бюджетных обследований являются обеспечение представительства всех социальных групп и учет всех источников дохода. Наиболее общим показателем уровня благосостояния населения являются денежные доходы, поступающие в семью в виде заработной платы, премий, единовременных выплат, гонораров, предпринимательского дохода или дохода от собственности, компенсационных выплат и дотаций. В совокупные доходы семьи включаются также натуральная оплата труда, доходы, полученные от реализации и потребления продукции личного подсобного хозяйства (садового участка, коллективного огорода). Для характеристики обеспеченности семей следует учитывать их накопления, а также валютные поступления. Возрастает значение анализа личного потребления.

Для изучения структуры рабочего времени работников разных категорий, особенно рабочих, а также для характеристики использования машин и оборудования используется *метод моментных наблюдений*. Этот метод состоит в регистрации вида затрат времени в определенные, заранее выбранные моменты. Предварительно составляется список всех возможных состояний или видов затрат времени. Подсчитывается доля отметок о каждом состоянии, и оценивается доверительный интервал доли времени, затраченного на тот или иной вид работы. Отбор моментов выборки может быть проведен либо по схеме механической выборки — через равные промежутки времени, либо по схеме случайной выборки с использованием таблицы случайных чисел. Необходимая численность моментов наблюдения рассчитывается как

$$n = \frac{0{,}25t^2}{\Delta^2}.$$

Если принять доверительную вероятность равной 0,954 и допустимую ошибку 0,005, т.е. 0,5%, то $n = \frac{0{,}25 \cdot 4}{0{,}000025} = 40\,000$ наблюдений. При продолжительности наблюдений 10 дней и охвате наблюдением 100 рабочих в день должно проводиться 40 наблюдений за каждым рабочим. Если продолжительность смены составляет 8 ч, то интервал между обходами должен составлять 12 мин [(8 ч · 60 мин) : 40 наблюдений]. За начальный момент можно принять момент начала смены плюс полинтервала: 7 ч 30 мин + 6 мин = 7 ч 36 мин, тогда второй обход будет проводиться в 7 ч 48 мин и т.д. По итогам моментного наблюдения рассчитываются доверительные интервалы для каждого i-го вида потерь рабочего времени:

$$p_i - \Delta_{p_i} \leqslant n_i \leqslant p_i + \Delta_{p_i}.$$

Результаты наблюдений используются для анализа потерь рабочего времени, статистической оценки напряженности труда рабочих, фактической загрузки оборудования и т.д.

Моментное наблюдение значительно менее трудоемко, чем, например, хронометраж или фотографии рабочего дня, а результаты (при надлежащей его организации) дает достаточно точные. Перед проведением наблюдения составляют общий план обследования, в том числе план расположения станков или рабочих мест, устанавливают маршрут, по которому должен следовать наблюдатель, определяют длину маршрута и время, необходимое на обход. Регистратор, проходя по маршруту, фиксирует в бланке конкретное действие конкретного рабочего или станка в данный момент времени («работа», «простой»). Удельный вес времени, в течение которого рабочий не работал, определяется отношением чисел отметок по элементу «простой» к общему числу отметок. Скажем, в течение 8 ч было зарегистрировано 56 моментов, среди которых 7 — по элементу «простой», 49 — по элементу «работа». В данном случае время работы составит 87,5% ($\frac{49}{56}100\%$), время простоев — 12,5% ($\frac{7}{56}100\%$).

Отбор моментов в выборочную совокупность осуществляется рядом способов. В силу необратимости времени он все-

гда является бесповторным. Но поскольку общее количество существующих моментов времени (генеральная совокупность) очень большое, в моментном наблюдении используют формулы ошибок повторного отбора.

Объем выборки для моментного наблюдения может быть рассчитан по формуле

$$n = \frac{t^2(1-k)100^2}{p^2 k},$$

где k — удельный вес изучаемого элемента в долях (обычно это коэффициент загруженности рабочих, или использования оборудования, определяемый примерно на основе отчетных данных);

p — заранее установленная относительная ошибка наблюдения в процентах.

Как правило, t в условиях стабильного производственного процесса принимается равным 2, нестабильного процесса — 3.

Пример. Определим количество необходимых наблюдений за работой станочников в условиях стабильного производственного процесса. На основе отчетных данных коэффициент загруженности рабочих (k) составляет 0,8, величина ошибки (p) принята равной 4%. Отсюда:

$$n = \frac{2(1-0{,}8)100^2}{4^2 \cdot 0{,}8} \cong 312 \text{ наблюдений.}$$

Если на участке имеется 52 рабочих места, то чтобы зафиксировать 312 наблюдений в течение 8-часового рабочего дня, один регистратор должен сделать 312 : 52 = 6 обходов. На один обход регистратор будет иметь 480 : 6 = 80 мин.

Допустим, после наблюдения установлено, что фактическая загруженность рабочих (количество отметок по элементу «работа») составила 243 наблюдения из 312. Следовательно, фактический коэффициент загруженности рабочих будет: $\frac{243}{312} = 0{,}78$, или 78% (а не 80%), потери рабочего времени — $\frac{69}{312} = 0{,}22$, или 22%, что в часах составляет соответственно 6,24 и 1,76.

Выборочный метод применяется в *аудиторской практике* при проверке бухгалтерских документов. При этом решаются две задачи: 1) оценка количества документов в данной фирме (предприятии, объединении и т.д.), в оформлении которых не соблюдались принятые правила; 2) оценка правильности указанных в документах сумм денежных средств. Первую задачу решают с помощью так называемой *атрибутивной выборки*, вторую — *монетарной выборки*. В первой выборке единицей отбора является учетный документ, во второй — денежная единица.

При организации атрибутивной выборки в качестве генеральной совокупности выступает вся совокупность расчетных документов фирмы за проверяемый период. Обычно она предварительно разбивается на однородные массивы — по характеру документов, по центрам ответственности, по географическому признаку, по временной последовательности, по интенсивности запросов на данный вид информации и т.д. Каждому документу присваивается числовая метка, и по таблице случайных чисел проводится отбор номеров в количестве, соответствующем объему выборки. Можно провести и механический отбор с шагом отбора, равным $N : n$, где N — объем генеральной совокупности, n — объем выборки. Обычно начинают отбор не с первого документа, а отступив полшага.

Объем атрибутивной выборки находится из соотношения

$$n = \frac{\text{Коэффициент надежности}}{\text{Максимально допустимая частота отклонений от стандартов оформления документов}} = \frac{R}{\Delta}.$$

Коэффициент надежности определяется по таблице распределения Пуассона, поскольку появление ошибки в оформлении расчетных документов относится к классу редких событий. При этом предполагаемая средняя частота ошибок закрепляется на определенном уровне, например 1, 1,5 или 2.

Если фактическая частота несоответствий в оформлении документов меньше максимально допустимой, то вычисляют коэффициент надежности как произведение объема выборки на величину фактической частоты несоответствий, после чего по таблице распределения Пуассона определяют вероятность, соответствующую рассчитанной величине коэффициента на-

дежности, чтобы убедиться, что доверительная вероятность результатов выборки достаточно высока.

Если фактически выявленная частота несоответствия превышает максимально допустимую величину, то обязательно проводят монетарную выборку.

При монетарной выборке генеральной совокупностью является сумма денежных средств, зафиксированных во всех проверяемых документах. В качестве единицы отбора выступает денежная единица (1 руб.), а единицей наблюдения является расчетный документ. Требуемая точность результатов задается как допустимая относительная сумма ошибки. Объем монетарной выборки рассчитывается как

$$n = \frac{\text{Коэффициент надежности}}{\text{Максимально возможная относительная сумма ошибок в документах}} = \frac{R}{\Delta/N} = \frac{RN}{\Delta}.$$

Например, если аудитор исходит из 1%-ного риска (при односторонней критической области из опасения, что суммарная ошибка будет не больше принятой величины), т.е. при 98%-ной доверительной вероятности наличия суммарной ошибки 50 000 руб., при объеме генеральной совокупности, равном 60 млн руб., то объем выборки

$$n = \frac{2{,}31 \cdot 60000000}{50000} = 2772 \text{ руб.}$$

Определяется шаг отбора, равный N: $n = 60\,000\,000 : 2772 = 21\,645$ руб. Все расчетные документы, в которых зафиксирована сумма, равная или превышающая величину шага отбора, обязательно попадут в выборку. Начало отбора устанавливается произвольно.

Пример. Рассмотрим записи по счету «Расчеты с покупателями» (табл. 7.5).

Приведенный пример показывает, что число отобранных документов может быть значительно меньше объема выборки по числу отбираемых денежных единиц. Если сумма операций многократно превышает шаг отбора, мы получаем несколько раз указание на необходимость проверки этой операции (в примере операция № 5 получила представительство в выборке шесть раз), и, наоборот, если сумма операции меньше шага отбора, она может не попасть в выборку (операция № 4). В целом чем крупнее операции по сравнению с шагом отбора, тем меньше будет совокупность отобранных документов — единиц наблюдения по сравнению с числом отобранных единиц.

Таблица 7.5

Формирование монетарной выборки, руб.
(в качестве начала отбора принято 25 000 руб., шаг отбора равен 21 645 руб.)

Номер операции	Сумма	Нарастающий итог	Отбираемая единица
1	22 000	22 000	—
2	10 000	32 000	25 000
3	18 500	50 500	46 645
4	10 275	60 775	—
5	126 850	187 625	68 290
			89 935
			111 580
			133 225
			154 870
			176 515
6	12 590	200 215	198 160
.	.	.	.
.	.	.	.
.	.	.	.

Решение вопросов по определению репрезентативности выборки и распространению ее результатов на генеральную совокупность зависит от того, были ли выявлены ошибки в выборке или нет. Это влияет на значение коэффициента надежности: сохранится оно или не сохранится. Исходя из этого проводится проверка соответствия фактической точности тому значению максимально допустимой суммарной величины ошибки, которое закладывалось при проектировании выборки. Если фактическая точность меньше или равна принятой, то выборка признается репрезентативной, если превышает ее, то применяются специальные методы оценки данных. Проверка проводится на основе соотношения

$$\text{Шаг отбора} = \frac{\text{Допустимый размер ошибки } (\Delta)}{\text{Коэффициент надежности } (R) \text{ с учетом фактического обнаружения ошибок}},$$

отсюда: $\Delta_{\text{факт}} = R \cdot$ шаг отбора.

Если при проверке отобранных документов ошибок не обнаружено, то с принятой доверительной вероятностью мы можем распространить результаты выборки на всю генеральную совокупность и считать, что итог по генеральной совокупности завышен не более чем на величину предельно допустимой ошибки. Если же обнаружена по крайней мере одна ошибка, то первоначальная гипотеза относительно отсутствия ошибок, которая закладывалась при планировании выборки, оказывается несостоятельной. В этом случае должны быть пересмотрены либо значение коэффициента надежности, либо величина предельно допустимой ошибки (точность), либо и то, и другое. Если ошибки выявлены в операциях, значение которых превышает величину шага отбора, то можно быть уверенным в отношении абсолютного размера ошибок в таких операциях, так как каждая из них проверялась полностью. В этом случае нужно решить вопрос о распространении абсолютного размера выявленных ошибок на операции, значение которых меньше шага отбора.

Все ошибки группируются в два класса: завышение суммы и ее занижение. Для всех операций, значение которых превышает шаг отбора, выявленная ошибка является точным размером завышения или занижения. Для операций, значение которых меньше шага отбора, размер выявленной ошибки относится к значению операции, и полученная относительная ошибка умножается на шаг отбора, т.е. распространяется на весь интервал (табл. 7.6).

После определения суммарного размера ожидаемой ошибки по всем интервалам выборки (т.е. шагам отбора) проводится сравнение с допустимым размером суммарной ошибки, и если рассчитанная суммарная ошибка превосходит допустимую величину, то, подставляя последнюю в формулу объема выборки, определяют, с каким коэффициентом надежности и соответственно с какой доверительной вероятностью могут гарантироваться результаты данного выборочного исследования:

$$R = \frac{n\Delta}{N}.$$

Таблица 7.6

Расчет суммарной ошибки на основе распространения результатов выборки, руб.

Характер и размер ошибки	Шаг отбора	Значение операции	Ожидаемый размер ошибки в интервале выборки
Завышение			
2500	21 645	126 850	2500
250	21 645	18 500	291,21 (установленный размер ошибки 1,35% распространен на весь интервал)
Итого 2750			2791,21
Занижение			
200	21 645	10 000	432,90
.	.	.	.
.	.	.	.
.	.	.	.

Как известно, в экономических исследованиях обычно принимают доверительную вероятность не ниже 90%.

Использование выборочного метода в работе аудитора резко повышает эффективность получения результатов и приводит к экономии финансовых и трудовых затрат.

Еще одним объектом применения выборочного метода является *малый бизнес*. В нашей стране работа по организации и проведению выборочного наблюдения малых предприятий включает следующие основные этапы.

1. Создание базы данных как основы выборки (базовой совокупности). Базовая совокупность (БС) включает список предприятий, определяемый рамками обследования малых предприятий, с набором показателей, полученных из единой генеральной совокупности (ГС). Начиная с 1998 г. для проведения выборочных обследований субъектов малого бизнеса формируется одна базовая совокупность единиц наблюдения (на основе генеральной совокупности объектов статистического наблюдения). Базовая совокупность создается раз в год и фиксируется по состоянию на 31 декабря предшествующего года. Базовая совокупность включает актуализированные в

части фактического основного вида деятельности признаки титульно-адресной части зарегистрированных в ЕГРПО малых предприятий, за исключением прекративших или приостановивших свою деятельность, и показатель выручки из бухгалтерской отчетности за год $t = 2$ (в 1999 г. за 1997 г. и т.д.). Основа выборки формируется из генеральной совокупности, зафиксированной по состоянию на 1 января 1999 г.

2. Формирование выборочной совокупности. Выборочная совокупность формируется на основе БГС методом стратифицированного случайного отбора с оптимизацией по Нейману. Выборочная совокупность формируется раз в год и фиксируется. При планировании выборок расслоение на региональном уровне (разделение на подсовокупности) базовой совокупности осуществляется по следующим признакам:

- ОКОНХ (на 62 страты);
- КФС (на 4 страты);
- ВЫРУЧКА (на 5 страт).

Показателем оптимизации является выручка. Если число единиц наблюдения в БГС невелико, то число страт по переменной «*выручка*» может быть уменьшено, и, наоборот, если в БГС значительное число единиц наблюдения (более 5000—7000), то число страт по переменной «*выручка*» может быть увеличено.

Объем выборки рассчитывается вычислительным комплексом автоматически при условии, что предельная ошибка по показателю «выручка» не должна превышать 5%-ный уровень.

3. Сбор и ввод (импорт) текущих данных отчетности по выборке. Собранные данные вводятся и контролируются средствами электронных версий формы №-ПМ с последующим вводом в вычислительный комплекс выборочного наблюдения.

4. Обработка полных неответов (восстановление пропусков). Практически при всех обследованиях предприятий имеются неответы респондентов опросного списка. Очень редко неответы бывают случайными. Систематические неответы могут вызвать смещение в оценках показателей в конкретном обследовании. При проведении статистических обследований различают два вида пропусков в данных: *полные неответы* — если в составе бланков форм отчетности с данными полностью отсутствуют результаты обследования по единице наблюдения. *Частичный неответ* или *пропуск* — при отсутствии

данных не в целом по единице наблюдения, а лишь по некоторым пунктам формуляра наблюдения. К частичным пропускам относят также ошибочные и некорректные ответы, которые могут быть внесены в бланк с данными в силу непонимания вопроса, неточности или просто невнимательности. Для обработки полных неответов респондентов совокупность неответивших предприятий должна быть разделена на три следующие группы:

- *первая* — предприятия, данные по которым восстанавливаться не будут. К ним относятся предприятия, ликвидированные или находящиеся в стадии ликвидации, так называемые спящие, т.е. приостановившие свою деятельность в силу различных причин;
- *вторая* — предприятия, о которых достоверно известно, что они, несмотря на отсутствие отчета, активны, ведут финансово-хозяйственную деятельность;
- *третья* — предприятия, по которым нет никаких данных и даже сведений, действующие они или нет.

К каждой группе полных неответов применяется свой метод коррекции и восстановления данных. Используются следующие методы восстановления пропусков:

- заполнение с пристрастным подбором;
- заполнение по предыдущему значению;
- заполнение без подбора;
- заполнение средними;
- заполнение с помощью регрессии;
- замена.

Заполнение с пристрастным подбором означает поиск данных, относящихся к единицам определенного типа.

Заполнение по предыдущему значению часто используется в современной практике. Но этот метод не рекомендуется применять при большом количестве пропусков, а также при наличии тенденции изменения показателя и значительном сроке со дня последней регистрации значения.

Заполнение безусловными средними. По имеющимся наблюдениям рассчитываются средние, и существующий пропуск заполняется средними значениями. Этот метод эффективен при однородности анализируемой совокупности и небольшом количестве пропусков.

Заполнение с помощью регрессии состоит в заполнении пропусков значениями, предсказываемыми регрессией пропущенных для данного объекта переменных на основе присутствующих. Регрессия вычисляется по объектам с полной информацией. Этот метод выдвигает ряд серьезных требований к данным: однородность, поскольку известно, что при использовании метода наименьших квадратов небольшое число грубых ошибок может весьма существенно исказить значение характеристики распределения; подчинение теоретическому нормальному распределению, что требует дополнительной обработки информации.

5. Досчет на вновь зарегистрированные предприятия. Записи о вновь зарегистрированных предприятиях добавляются к выборочной совокупности, и коэффициент увеличения численности используется как коэффициент досчета по всем показателям.

6. Распространение результатов выборочного наблюдения на генеральную совокупность проводится по методике, рассмотренной выше.

7. Анализ и экспертная корректировка полученных результатов. За качество передаваемой на федеральный уровень информации отвечает соответствующая территория (субъект РФ или федеральный округ). Достоверность отчетности зависит только от квалификации исполнителя и желания добросовестно сделать свою работу.

Решению проблем, связанных прежде всего с проблемами организации и проведения выборочных обследований малых предприятий на региональном уровне, посвящена разработка подпроекта Программы TACIS «Статистика-3». Особое внимание уделялось вопросам подготовки анкеты выборочного наблюдения, составу и структуре содержащихся в ней показателей, а также концепциям формирования выборки на региональном уровне.

Большая проблема для российской статистики состоит в выявлении и обработке данных нетипичных единиц наблюдения. Несмотря на достаточно эффективный план выборки проводимого обследования, при детальном анализе данных на региональном и федеральном уровнях неоднократно выявлялись единицы, включение (или исключение) которых в выборочную совокупность сильно влияет на итоговое значение получаемых оценок показателей и в конечном счете отражается на качестве результатов обследования.

Нетипичные единицы определяются как:

- имеющие экстремальные значения показателей;
- влияющие на конечную оценку из-за своего большого выборочного веса;
- имеющие сложную структуру или находящиеся в процессе структурной перестройки.

Для выявления и оценки влияния нетипичных единиц на конечные значения показателей обследования используются следующие методы:

- *графический метод* — прост в исполнении, но может применяться только в случае небольших объемов совокупности единиц наблюдения;
- *квартильный метод* — удобный и широко используемый на практике. Суть его заключается в построении с помощью медианы межквартильных рангов границ предельно допустимого интервала. Единицы, значения признаков которых попадают за рамки этого интервала, являются нетипичными;
- *агрегированный контроль* — этап обработки индивидуальных данных, проводимый перед распространением результатов наблюдения на исследуемую совокупность. В качестве показателей для агрегированного контроля в обследовании малых предприятий по форме №-ПМ выбираются: выручка, численность занятых, фонд заработной платы, а также выручка/численность, фонд заработной платы/численность. Если предприятие не прошло агрегированный контроль, то оно заносится в перечень для представления руководителю обследования, исправляющему или подтверждающему данные по этому предприятию. Далее осуществляется индивидуальный контроль только за выявленными единицами.

Агрегированный контроль можно назвать макроэкономическим, так как в процессе его проведения используются соотношения показателей не на индивидуальном уровне, а на уровне страны, отрасли.

При больших объемах первичной информации методика выявления нетипичных единиц с применением пакета SPSS, разработанная Госкомстатом России, может служить допол-

нительным контролем при разработке итогов обследований малых предприятий.

Выборочный метод широко используется при проведении *конъюнктурных опросов*. Конъюнктурные опросы рекомендуется проводить по постоянной выборке, т.е. по *панели предприятий*. Это обеспечивает существенные преимущества при организации опросов и анализе результатов.

Достоинства панельной организации опросов.

Во-первых, регулярное получение ответов от одной и той же совокупности предприятий создает уникальную возможность экономического анализа на микроуровне.

Во-вторых, при разумной и дальновидной организации хранения и накопления результатов панельных опросов появляется возможность многократного и всестороннего использования результатов опросов. При этом аналитические результаты могут быть получены без проведения новых опросов, а только за счет применения новых методов или моделей к уже накопленным данным. Новые опросы на той же панели могут в этом случае проводиться для расширения уже существующих первичных данных.

В-третьих, регулярный (ежемесячный или ежеквартальный) характер бизнес-обследований позволяет организаторам при необходимости регулярно совершенствовать вопросы анкеты и получать таким образом все более точные данные об исследуемых явлениях.

В-четвертых, создание панели и накопление панельных данных позволяют использовать специфические статистические методы и эконометрические модели, не применимые к другим типам данных. Эти методы и модели способны обеспечить получение принципиально новых результатов.

В настоящее время на регулярной основе проводятся обследования предпринимательских намерений в промышленности, строительстве, сельском хозяйстве, оптовой торговле, а также в банковском и страховом секторах и в инновационной сфере.

Обследования базовых предприятий промышленности проводятся ежемесячно; по промышленности в целом — ежеквартально; строительных организаций, оптовой торговли и в инновационной сфере — ежеквартально; в банковском и страховом секторах — два раза в год.

Выборочное наблюдение широко используется при изучении качества готовой продукции. Отбор готовых изделий для установления их качества проводится главным образом механически (5-е, 10-е, 15-е изделие и т.д.). Если изделия в таре, то в большинстве случаев осуществляется серийный отбор (единица отбора = единице тары). Это так называемый приемочный или последующий контроль, основанный на проверке качества уже выработанных изделий; он не в состоянии предупредить появление брака.

Большое распространение получил непрерывный текущий статистический контроль за качеством изготовляемой продукции, осуществляемый в форме отбора проб в ходе производственного процесса непосредственно у рабочих мест. Такой контроль обеспечивает систематическое наблюдение не только за качеством продукции, но и за самим производственным процессом. Текущий контроль в ходе отбора и анализа проб позволяет своевременно обнаружить неполадки в работе, сигнализировать о них и тем самым предупредить возникновение брака.

Значительной сферой применения выборочного наблюдения являются маркетинговые исследования, проводимые с целью оценки мощности рынков товаров и услуг, определения специфических сегментов рынка.

РЕЗЮМЕ

Выборочное наблюдение проводится с целью повышения точности и оперативности данных, экономии материальных, трудовых и финансовых ресурсов.

Для того чтобы по выборке можно было делать вывод о свойствах генеральной совокупности, выборка должна быть репрезентативной.

Репрезентативность выборки может быть обеспечена объективным отбором данных. Используют три способа отбора: случайный, механический, сочетание первого и второго способов.

Если отбор проводится из генеральной совокупности, предварительно разделенной на типы (районы, слои или страты), то такая выборка называется типической (районированной, расслоенной или стратифицированной).

Единицей отбора может быть единица наблюдения или группа единиц. В последнем случае выборка называется серийной или гнездовой. В социально-экономических исследованиях используется схема бесповторной выборки.

Ошибки выборочного наблюдения подразделяются на случайные и неслучайные. Случайные ошибки подчиняются вероятностным законам. К случайным относится ошибка выборки, называемая ошибкой репрезентативности. Рассчитываются ошибки выборки для выборочных средних и выборочных относительных величин.

На величину ошибки выборки влияет вид выборки: если районы существенно отличаются друг от друга, то ошибка районированной выборки будет меньше, чем нерайонированной выборки; применение гнездовой выборки при прочих равных условиях приводит к увеличению ошибки выборки. На практике часто используют сочетание районированной выборки с гнездовым отбором.

Применение выборочного метода связано с решением трех задач:

- определение объема выборки, обеспечивающего требуемую точность результатов с принятой вероятностью;
- расчет предельной ошибки репрезентативности, гарантированный с принятой вероятностью, и сравнение его с величиной допустимой погрешности;
- определение вероятности того, что ошибка выборки не превысит допустимой погрешности.

Первая задача связана с распространением данных выборки на генеральную совокупность. На основе выборочных характеристик даются интервальные оценки генеральных параметров. Могут быть получены и оценки значения подсчетов в генеральной совокупности.

Определенные особенности имеют организация и проведение малых выборок (при $n \leqslant 30$ единиц).

Выборочный метод все шире применяется как в официальной статистике, так и в научных исследованиях, и в бизнесе.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ: Пер. с англ. / Под ред. Г. П. Башарина. — М.: Мир, 1982.
2. *Бокун Н. Ч., Чернышова Н. М.* Методы выборочных обследований. — Минск: Министерство статистики и анализа Республики Беларусь. НИИ статистики, 1997.
3. *Головач А. В., Ерина А. М., Трофимов В. П.* Критерии математической статистики в экономических исследованиях. — М.: Статистика, 1973.
4. *Джессен Р.* Методы статистических обследований: Пер. с англ. / Под ред. и с предисл. Е. М. Четыркина. — М.: Финансы и статистика, 1985.
5. *Дружинин Н. К.* Математическая статистика в экономике. — М.: Статистика, 1971.
6. *Информатика* в статистике: словарь-справочник. — М.: Финансы и статистика, 1994.
7. *Йейтс Ф.* Выборочный метод в переписях и обследованиях. — М.: Статистика, 1965.
8. *Кокрен У.* Методы выборочного исследования: Пер. с англ. / Под ред. А. Г. Волкова. — М.: Статистика, 1976.
9. *Паниотто В. И.* Качество социологической информации (Методы оценки и процедуры обеспечения). — Киев: Наукова думка, 1986.
10. *Фишер Р. А.* Статистические методы для исследователей: Пер. с англ. — М.: Госстатиздат, 1958.

Глава

8

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

8.1. Общие понятия

В гл. 7 оценка генерального параметра была получена на основе выборочного показателя с учетом ошибки репрезентативности. В отношении свойств генеральной совокупности могут выдвигаться некоторые гипотезы о величине средней, дисперсии, характере распределения, форме и тесноте связи между переменными. Проверка гипотезы осуществляется на основе выявления согласованности эмпирических данных с гипотетическими (теоретическими). Если расхождение между сравниваемыми величинами не выходит за пределы случайных ошибок, гипотезу принимают. При этом не делается никаких заключений о правильности самой гипотезы, речь идет лишь о согласованности сравниваемых данных. Основой проверки статистических гипотез являются данные случайных выборок. При этом безразлично, оцениваются ли гипотезы в отношении реальной или гипотетической генеральной совокупности. Последнее открывает путь применения этого метода за пределами собственно выборки: при анализе результатов эксперимента, данных сплошного наблюдения, но малой численности. В этом случае рекомендуется проверить, не вызвана ли установленная закономерность стечением случайных обстоятельств, насколько она характерна для того комплекса условий, в которых находится изучаемая совокупность.

Особенно часто процедура проверки статистических гипотез проводится для оценки существенности расхождений сводных характеристик отдельных совокупностей (групп): средних, относительных величин. Такого рода задачи, как правило,

возникают в социальной статистике. Трудоемкость статистико-социологических исследований приводит к тому, что почти все они строятся на несплошном учете. Поэтому проблема доказательности выводов в социальной статистике стоит особенно остро. Применяя процедуру проверки статистических гипотез, следует помнить, что она может гарантировать результаты с определенной вероятностью лишь по «беспристрастным» выборкам, на основе объективных данных.

Статистической гипотезой называется предположение о свойстве генеральной совокупности, которое можно проверить, опираясь на данные выборки. Обозначается гипотеза буквой H (лат. hypothesis). Так, может быть выдвинута гипотеза о том, что средняя в генеральной совокупности равна некоторой величине $H: \mu = a$, или о том, что генеральная средняя больше некоторой величины: $H: \mu > b$.

Различают *простые* и *сложные* гипотезы. Гипотеза называется простой, если она однозначно характеризует параметр распределения случайной величины. Например, $H: \mu = a$. Сложная гипотеза состоит из конечного или бесконечного числа простых гипотез, при этом указывается некоторая область вероятных значений параметра. Например, $H: \mu > b$. Эта гипотеза состоит из множества простых гипотез: $H: \mu = c$, где c — любое число, большее b.

Гипотезы о параметрах генеральной совокупности называются *параметрическими*, о распределениях — *непараметрическими*.

Гипотеза о том, что две совокупности, сравниваемые по одному или нескольким признакам, не отличаются, называется *нулевой гипотезой* (или нуль-гипотезой). Она обозначается H_0. При этом предполагается, что действительное различие сравниваемых величин равно нулю, а выявленное по данным отличие от нуля носит случайный характер. Например, $H_0: \mu_1 = \mu_2$ и т.д.

Нулевая гипотеза отвергается тогда, когда по выборке получается результат, который при истинности выдвинутой нулевой гипотезы маловероятен. Границей невозможного или маловероятного обычно считают $\alpha = 0{,}05$, т.е. 5%, или 0,01, 0,001. Если ориентироваться на правило «трех сигм», то вероятность ошибки α должна быть равна 0,0027. Однако для этого уровня вероятности ошибки значения критериев редко табу-

лируются: как правило, значения критериев в статистико-математических таблицах рассчитаны для вероятностей ошибки 0,05; 0,01; 0,001.

Статистическим критерием называют определенное правило, устанавливающее условия, при которых проверяемую нулевую гипотезу следует либо отклонить, либо не отклонить. Критерий проверки статистической гипотезы определяет, противоречит ли выдвинутая гипотеза фактическим данным или нет.

Проверка статистических гипотез складывается из следующих этапов:

- формулируется в виде статистической гипотезы задача исследования;
- выбирается статистическая характеристика гипотезы;
- выбираются испытуемая и альтернативная гипотезы на основе анализа возможных ошибочных решений и их последствий;
- определяются область допустимых значений, критическая область, а также критическое значение статистического критерия (t, F, χ^2) по соответствующей таблице;
- вычисляется фактическое значение статистического критерия;
- проверяется испытуемая гипотеза на основе сравнения фактического и критического значений критерия, и в зависимости от результатов проверки гипотеза либо отклоняется, либо не отклоняется.

При проверке гипотез по одному из критериев возможны два ошибочных решения:

1) неправильное отклонение нулевой гипотезы: ошибка 1-го рода;

2) неправильное принятие нулевой гипотезы; ошибка 2-го рода.

В то время как фактически нулевая гипотеза верна (1) и ненулевая гипотеза не верна (2), принимают два ошибочных решения: 1) нулевая гипотеза отклоняется и принимается альтернативная гипотеза; 2) нулевая гипотеза не отклоняется. Возможные решения представлены в табл. 8.1.

Если, например, установлено, что новое минеральное удобрение лучше, хотя на самом деле его действие не отличается от старого, то это ошибка 1-го рода. Если мы решили,

Таблица 8.1

Возможные решения при проверке гипотез

Решение по критерию	Фактически	
	H_0 верна	H_0 не верна
H_0 отклоняется	Ошибка 1-го рода	Правильное решение
H_0 не отклоняется	Правильное решение	Ошибка 2-го рода

что оба вида удобрений одинаковы, то допущена ошибка 2-го рода.

Вероятности, соответствующие неверным решениям, называются риском 1 и риском 2. *Риск 1* равен вероятности ошибки α (*уровню значимости*), *риск 2* равен вероятности ошибки β. Поскольку α всегда больше нуля, то всегда есть риск ошибки β. При заданных α и объеме выборки n значение β будет тем больше, чем меньше принятое α. Если n велико, то α и β могут быть сколь угодно малыми, т.е. решения будут более обоснованными. При малом объеме выборки и малом α возможность установить фактически существующие различия мала.

Обычно задают значение α и пытаются сделать β возможно малым. Вероятность $1 - \beta$ называется *мощностью критерия*: чем она больше, тем меньше вероятность ошибки 2-го рода.

Альтернативная гипотеза H_1 может быть сформулирована по-разному в зависимости о того, какие отклонения от гипотетической величины нас особенно беспокоят: положительные, отрицательные либо и те, и другие. Соответственно альтернативные гипотезы могут быть записаны как

$$H_1 : \mu > a;\ H_1 : \mu < a;\ H_1 : \mu \neq a.$$

От того как формулируется альтернативная гипотеза, зависят границы критической области и области допустимых значений.

Критической областью называется область, попадание значения статистического критерия в которую приводит к отклонению H_0. Вероятность попадания значения критерия в эту область равна принятому уровню значимости.

Область допустимых значений дополняет критическую область. Если значение критерия попадает в область допустимых значений, это свидетельствует о том, что выдвинутая гипотеза H_0 не противоречит фактическим данным (H_0 не отклоняется).

Точки, разделяющие критическую область и область допустимых значений, называются *критическими точками* или *границами критической области*. В зависимости от формулировки альтернативной гипотезы критическая область может быть *двусторонняя* (рис. 8.1, *а*) или *односторонняя* (рис. 8.1, *б*) — левосторонняя либо правосторонняя.

Если вычисляемое значение критерия попадает в критическую область, нулевая гипотеза отклоняется, поскольку она противоречит фактическим данным.

8.2. Проверка гипотезы о законе распределения

Одна из важнейших задач анализа вариационных рядов заключается в выявлении закономерности распределения и определении ее характера. Основной путь в выявлении закономерности распределения — построение вариационных рядов для достаточно больших совокупностей. Важное значение для выявления закономерности распределения имеет правильное построение самого вариационного ряда: выбор числа групп и размера интервала варьирующего признака.

Когда мы говорим о характере, типе закономерности распределения, имеем в виду отражение в нем общих условий вариации. При этом речь всегда идет о распределениях качественно однородных явлений. Общие условия, определяющие тип закономерности распределения, познаются анализом сущности явления, тех его свойств, которые определяют вариацию изучаемого признака. Следовательно, должна быть выдвинута какая-то научная гипотеза, обосновывающая тип теоретической кривой распределения.

Под теоретической кривой распределения понимается графическое изображение ряда в виде непрерывной линии изменения частот в вариационном ряду, функционально связанного с изменением вариантов (значений признака). Теоретическое распределение может быть выражено аналитически — формулой, которая связывает частоты вариационного ряда и соответствующие значения признака. Такие алгебраические формулы носят название ***законов распределения***.

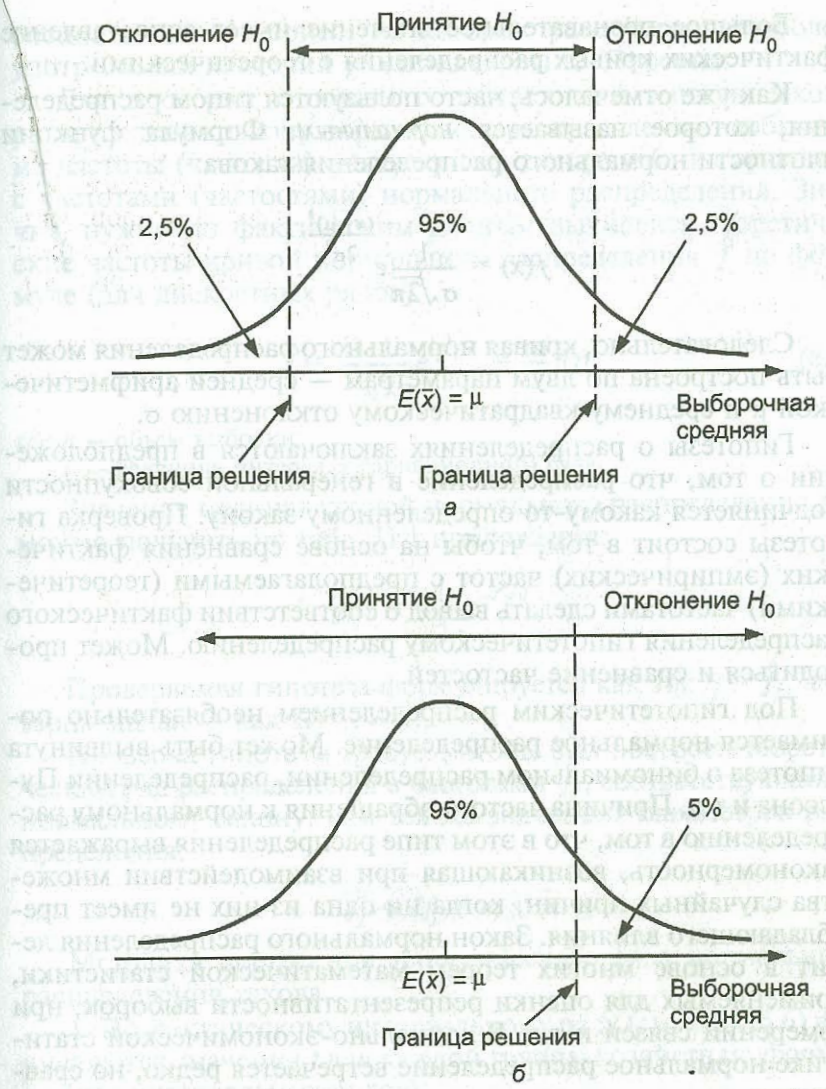


Рис. 8.1. 5%-ная проверка:
а — двусторонняя; *б* — односторонняя

Большое познавательное значение имеет сопоставление фактических кривых распределения с теоретическими.

Как уже отмечалось, часто пользуются типом распределения, которое называется *нормальным*. Формула функции плотности нормального распределения такова:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}.$$

Следовательно, кривая нормального распределения может быть построена по двум параметрам — средней арифметической μ и среднему квадратическому отклонению σ.

Гипотезы о распределениях заключаются в предположении о том, что распределение в генеральной совокупности подчиняется какому-то определенному закону. Проверка гипотезы состоит в том, чтобы на основе сравнения фактических (эмпирических) частот с предполагаемыми (теоретическими) частотами сделать вывод о соответствии фактического распределения гипотетическому распределению. Может проводиться и сравнение частостей.

Под гипотетическим распределением необязательно понимается нормальное распределение. Может быть выдвинута гипотеза о биномиальном распределении, распределении Пуассона и т.д. Причина частого обращения к нормальному распределению в том, что в этом типе распределения выражается закономерность, возникающая при взаимодействии множества случайных причин, когда ни одна из них не имеет преобладающего влияния. Закон нормального распределения лежит в основе многих теорем математической статистики, применяемых для оценки репрезентативности выборок, при измерении связей и т.д. В социально-экономической статистике нормальное распределение встречается редко, но сравнение с ним важно для выяснения степени и характера отклонения от него фактического распределения.

В гл. 5 отмечалось, что близость средней арифметической величины, медианы и моды указывает на вероятное соответствие изучаемого распределения нормальному закону. Но более полная и точная проверка соответствия распределения гипотезе о нормальном законе проводится с использованием

специальных критериев, из которых рассмотрим наиболее употребимый критерий χ^2 (хи-квадрат) К. Пирсона.

Для проверки гипотезы о соответствии эмпирического распределения закону нормального распределения необходимо частоты (частости) фактического распределения сравнить с частотами (частостями) нормального распределения. Значит, нужно по фактическим данным вычислить теоретические частоты кривой нормального распределения $\hat{f}$ по формуле (для дискретных рядов)

$$\hat{f} = \frac{ni}{s}\frac{1}{\sqrt{2\pi}} e^{-t^2/2} = \frac{ni}{s} f(t), \qquad (8.1)$$

где n — объем выборки;
i — величина интервала вариационного ряда.

Значение ординат кривой нормального распределения $f(t)$ можно получить по табл. П.1 приложения:

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Проверяемая гипотеза формулируется как $H_0: f_j = \hat{f}_j$, альтернативная — как $H_1: f_j \neq \hat{f}_j$.

Проверка гипотезы требует, чтобы был построен теоретический ряд распределения с частотами $\hat{f}_j$, соответствующими нормальному закону, при тех же значениях параметров распределения:

$$\Sigma \hat{f}_j = \Sigma f_j = n;\ s;\ \bar{x}.$$

Методика построения теоретического ряда нормального распределения такова.

1. По фактическому интервальному ряду (см. табл. 5.6) вычисляются значения t для каждой группы хозяйств по формуле (для интервальных рядов)

$$t_j = \frac{x_j - x}{s} \quad \text{— для начала и конца интервала.}$$

2. Вычисляется вероятность попадания единицы наблюдения в данный интервал при выполнении гипотезы о нормальном законе:

$$P_j = \frac{F(t_j) - F(t_{j+1})}{2},$$

где $|t_j| > |t_{j+1}|$.

3. Определяется теоретическая частота в данной группе, равная произведению объема совокупности на вероятность попадания в данный интервал:

$$\hat{f}_j = \sum_{j=1}^{k} f_j P_j = n \cdot P_j.$$

4. Находится значение критерия χ^2 по формуле

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j}, \qquad (8.2)$$

где k — число категорий ряда распределения;

f_j — частота эмпирического распределения;

$\hat{f}_j$ — частота теоретического распределения;

j — номер категории.

При расчете χ^2 частоты можно заменить частостями:

$$\chi^2 = n \sum_{j=1}^{k} \frac{(p_j - \pi_j)^2}{\pi_j}, \qquad (8.3)$$

где p_j — частости эмпирического распределения;

π_j — вероятности теоретического распределения.

При этом согласно Ф. Йейтсу группы с теоретическими частотами менее пяти принято объединять, что снижает влияние случайных ошибок [2].

Если все эмпирические частоты равны соответствующим теоретическим частотам, то χ^2 равно нулю. Очевидно, что чем больше отличаются эмпирические и теоретические частоты, тем χ^2 больше; если расхождение несущественно, то χ^2 должно быть малым. Имеются специальные таблицы критических значений χ^2 при 5%-ном и 1%-ном уровнях значимости (см. приложение). Критические значения зависят от числа степеней свободы (*d.f.*) и уровня значимости α.

Число степеней свободы рассчитывается так: если эмпирический ряд распределения имеет k категорий, то k эмпирических частот $f_1, f_2, ..., f_k$ должны быть связаны следующим соотношением: $\sum_{1}^{k} f_j = n$. Если параметры теоретического распределения известны, то только $k - 1$ частот могут принимать произвольные значения, т.е. свободно варьировать, а последняя частота может быть найдена из указанного соотношения. Поэтому говорят, что система из k частот благодаря наличию одной связи теряет одну степень свободы и имеет только $k - 1$ степеней свободы. Кроме того, если при нахождении теоретических частот p параметров теоретического распределения неизвестны, то они должны быть найдены по данным эмпирического ряда. Это накладывает на эмпирические частоты еще p связей, благодаря чему система теряет p степеней свободы. Таким образом, число свободно варьируемых частот (а значит, и число степеней свободы) становится равным:

$$d.f. = (k - 1) - p = k - (p + 1). \qquad (8.4)$$

Полученное значение критерия χ^2 сравнивается с табличным при числе степеней свободы, равном числу групп (с условием Ф. Йейтса), за минусом трех — по числу фиксированных параметров в формуле нормального закона распределения и с учетом равенства сумм теоретических и фактических частот (табл. П.3 приложения).

В первой графе этой таблицы дано число степеней свободы, а в заголовках граф — уровни значимости. Если фактическое значение χ^2 превышает табличное при том же числе степеней свободы, то вероятность соответствия распределения нормальному закону меньше указанной. Результаты расчета χ^2 по данным табл. 5.6 (см. гл.5) приведены в табл. 8.2 при $\bar{x} = 30{,}3$; $s = 8{,}44$.

Сумма теоретических частот нормального распределения оказалась меньше суммы фактических частот, так как нормальный закон не ограничен рамками фактических минимума и максимума.

Число групп после объединения малочисленных составило семь. Критическое значение χ^2 по табл. П.4 приложения при $7 - 3 = 4$ степеням свободы и значимости 0,05 составляет

Таблица 8.2

Проверка соответствия распределения хозяйств по урожайности зерновых культур нормальному закону

Группа хозяйств	f_j	t_j	t_{j+i}	P_j	$\hat{f}_j$	$(f_j - \hat{f}_j)^2 / \hat{f}_j^2$
1	6	−2,41	−1,81	0,0235	3	0,071
2	9	−1,81	−1,22	0,0798	11	
3	20	−1,22	−0,63	0,1531	22	0,182
4	41	−0,63	−0,04	0,2197	32	2,531
5	26	−0,04	0,56	0,2282	33	1,485
6	21	0,56	1,15	0,1627	23	0,174
7	14	1,15	1,74	0,0842	12	0,333
8	5	1,74	2,33	0,0310	4	0,200
9	1	2,33	2,93	0,0082	1	
Σ	143	×	×	0,9904	141	4,976

9,49. Значит, вероятность расхождения распределения с нормальным меньше 0,05 и вероятность соответствия его нормальному закону больше 0,95. Табличное значение χ^2 на уровне значимости 0,1 равно 7,78, что также больше фактического.

Ясно, что гипотеза о соответствии распределения хозяйств по урожайности нормальному закону не может быть отклонена.

Какое практическое значение может иметь проверка гипотезы? Во-первых, соответствие нормальному закону позволяет прогнозировать, какое число хозяйств (или доля совокупности) попадает в тот или иной интервал значений признака. Во-вторых, нормальное распределение возникает при действии на вариацию изучаемого показателя множества независимых факторов. Из этого следует, что нельзя существенно снизить вариацию урожайности, воздействуя лишь на один-два управляемых фактора, скажем удобрения или энергозатраты.

С помощью критерия χ^2 можно проверять не только гипотезу о согласии эмпирического распределения с нормальным законом, но и с любым другим известным законом распределения — равномерным распределением, распределением Пуассона и т.д.

Например, суд рассматривает жалобу посетителей казино на то, что игральная кость, которой там пользуются, по их

мнению, фальшива, некоторые числа очков, якобы, выпадают чаще, чем другие, и этим пользуются крупье, обирающие игроков.

Суд назначает экспертизу игральной кости: эксперт делает 600 бросков и записывает число выпавших единиц, двоек, троек и т.д. Полученное эмпирическое распределение сравнивается с теоретическим, т.е. равномерным: в правильной кости вероятность выпадения каждого числа очков должна быть равна 1/6, при 600 бросках это даст по 100 выпадений каждого числа очков. С помощью критерия χ^2 проверяется нулевая гипотеза о том, что различия эмпирического и теоретического распределений случайны, т.е. не являются систематическим результатом фальсификации формы кости или положения центра тяжести в ней; H_0: $f_{\text{факт}} = f_{\text{теор}}$. Результаты испытания и расчет χ^2 приводятся в табл. 8.3.

Табличное значение χ^2 при уровне значимости 0,05 (это вероятность ошибочного отклонения нулевой гипотезы при условии, что она верна) и при 6 − 2 = 4 степенях свободы (фиксированы два параметра: сумма числа бросков 600 и вероятность каждого числа очков — 1/6) составляет 9,49. Вычисленное значение χ^2 = 5,2, что значительно ниже табличного. Следовательно, нулевая гипотеза не отклоняется: распределение бросков по числу выпавших очков нельзя считать неравномерным. Обвинение игроками служащих казино не подтверждено достаточно надежно, но и не доказано то, что

Таблица 8.3

Результаты испытания игральной кости

Число очков	Количество выпадений, $f_{\text{факт}}$	$f_{\text{теор}}$	$f_{\text{факт}} - f_{\text{теор}}$	$\frac{(f_{\text{факт}} - f_{\text{теор}})^2}{f_{\text{теор}}}$
1	101	100	1	0,01
2	86	100	−14	1,96
3	107	100	7	0,49
4	94	100	−6	0,36
5	97	100	−3	0,09
6	115	100	15	2,25
Итого	600	600	0	5,16

кость правильная. Можно назначить более дорогую экспертизу — сделать 100 000 бросков кости, но можно и согласиться, что вероятность ошибочного признания правильности кости мала — всего 5%, и отклонить обвинение.

Распределение Пуассона описывает вероятность редких событий: гибель от укуса животного или удара копытом; в экономике это вероятность ошибок в финансовых документах или сверхкрупной взятки и т.п.

Если в схеме Бернулли p — малая величина, то вероятность $p_{n,\,m}$ можно найти по приближенной формуле

$$p_{n,\,m} = c_n^m p^m q^{n-m} \approx \frac{\lambda^m}{m!} e^{-\lambda} = p_m(\lambda), \qquad (8.5)$$

где $\lambda = np$ — среднее число появлений события A в n испытаниях;
m — число появлений события A в n независимых испытаниях.

Придавая m целые неотрицательные значения $m = 0, 1, 2, \ldots, n$ можно записать ряд распределения вероятностей, вычисленных по формуле (8.5), которая называется законом распределения Пуассона (табл. 8.4).

Формулой (8.5) можно пользоваться, когда $p \approx 0{,}1$ и $npq \leqslant 9$. Распределение Пуассона приведено в табл. П.8 приложения. Очевидно, что оно является предельным случаем распределения вероятностей в схеме Бернулли при $p \approx 0{,}1$, $np = \lambda$. Математическое ожидание и дисперсия случайной величины, распределенной по закону Пуассона, совпадают и равны параметру λ, который определяет этот закон:

$$M(x) = D(x) = \lambda.$$

Проверка гипотезы о том, что выборка извлечена из генеральной совокупности, имеющей распределение Пуассона, проводится следующим образом.

Таблица 8.4

Распределение вероятностей в соответствии с законом Пуассона

m	0	1	2	...	m	...	n
$p_{n,\,m}$	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2 e^{-\lambda}}{m!}$	...	$\frac{\lambda^m e^{-\lambda}}{m!}$	...	$\frac{\lambda^n e^{-\lambda}}{n!}$

1. По заданному дискретному вариационному ряду рассчитывают выборочную среднюю $\bar{x}$. Ее значение используется в качестве оценки параметра λ распределения Пуассона.

2. Вычисляются вероятности (частости) p_i:

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda},$$

$$i = 0, 1, \ldots, k-1;$$

$$p_k = p(x \geqslant k) = \sum_{i=k}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda}.$$

3. Умножая найденные значения вероятностей на объем выборки, рассчитываются теоретические частоты распределения:

$$np_i = \hat{f}_i.$$

4. Рассчитывается значение критерия согласия Пирсона хи-квадрат.

Находится критическое значение критерия хи-квадрат при заданном уровне значимости и числе степеней свободы: $d.f. = k - 1 - r$, где $r = \begin{cases} 1 \\ 0 \end{cases}$ в зависимости от того, оценивался ли параметр λ по выборке.

5. Сравниваются рассчитанное (фактическое) и критическое (табличное) значения критерия хи-квадрат и делаются выводы:

- если $\chi^2_{\text{факт}} < \chi^2_{\text{табл}}$, то нет оснований отклонять гипотезу о распределении случайной величины по закону Пуассона;
- если $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, то гипотеза отклоняется.

Пример. Для проведения внутреннего контроля качества оформления платежных требований в случайном порядке были выбраны 100 документов. В табл. 8.5 содержатся результаты проверки документов. Среднее количество ошибок составило $\bar{x} = 0{,}39$.

Значение $\chi^2_{\text{факт}} = 16{,}88$. Число степеней свободы составляет: $d.f. = 5 - 1 = 4$. Табличные значения критерия хи-квадрат оказались меньше фактического значения:

$$\chi^2_{d.f.\,=\,4;\ \alpha\,=\,0{,}05} = 9{,}499; \qquad \chi^2_{d.f.\,=\,4;\ \alpha\,=\,0{,}01} = 13{,}28.$$

Таблица 8.5

Результаты выборочного контроля платежных требований

Количество ошибок, x_i	Число проверенных документов, f_i	$p_i = \frac{0{,}39^{x_i} y^{-0{,}39}}{x_i!}$	$\hat{f}_i$	$\frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$
0	75	0,6761	67,6	0,8101
1	16	0,2622	26,2	3,9710
2	5	0,0505	5,1	0,0020
3	3	0,0105	1,0	4,000
4	1	0,0007	0,1	8,100
Итого	100	1,000	100	16,8831

Так как $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, гипотеза о распределении Пуассона отклоняется.

Пример. Проверкой установлено, что среди партии гаек на складе 0,45% имеют брак. Какова вероятность того, что при случайном отборе 200 гаек обнаружится 8 бракованных? Известно, что вероятность появления брака равна $p = 0{,}0045$, $\lambda = np = 9$, так как $n = 2000$.

Отсюда: $p_{2000,8} = p_8(9) = \frac{e^{-9} 9^8}{8!} = 0{,}1318$.

Тот же результат мы получим по таблице значений функции Пуассона (табл. П.8 приложения).

Критерий Колмогорова—Смирнова

Проверку гипотезы о законе распределения можно проводить с помощью критерия Колмогорова—Смирнова. Это альтернатива критерию хи-квадрат. Применение этого критерия не требует расчета ожидаемых частот и может использоваться для малых выборок. Данные должны представлять случайную выборку, переменные должны быть измерены по крайней мере на порядковой шкале; должна быть сформулирована гипотеза о распределении генеральной совокупности. Нулевая гипотеза состоит в том, что выборка взята из специфицированной генеральной совокупности. Альтернативная гипотеза заключается в утверждении обратного.

Применение критерия Колмогорова—Смирнова основано на кумулятивных частостях (вероятностях): наблюдаемых и ожидаемых, т.е. найденных в предположении, что нулевая гипотеза верна.

Тестовая статистика — это максимальная абсолютная разность наблюдаемой и ожидаемой частостей. Наблюдаемые данные ранжируются от минимальной величины до максимальной. В случае больших выборок могут использоваться сгруппированные данные (ряд распределения).

Тестовая статистика вычисляется по формуле

$$D = \max |F_i - E_i|, \qquad (8.6)$$

где F_i — наблюдаемая кумулятивная частость для i-го значения (или интервала);

E_i — ожидаемая кумулятивная частость для i-го значения (или интервала).

Если D больше критического значения, взятого из табл. П.9 приложения для объема выборки n и уровня значимости α, нулевая гипотеза должна быть отклонена. В противном случае нулевая гипотеза не может быть отклонена.

Пример. Получена случайная выборка данных о среднем дневном заработке, руб./день, для пяти работников: 288, 231, 249, 146, 291.

Можно ли считать на 10%-ном уровне значимости, что выборка проведена из нормально распределенной совокупности со средней величиной $\mu = 200$ руб./день и $\delta = 50$ руб./день?

Очевидно, что выборка слишком мала и нельзя применить критерий хи-квадрат. Ответ может быть дан только с помощью критерия Колмогорова—Смирнова. Нулевая гипотеза формулируется как

$$H_0: \mu = 200, \delta = 50;$$

$$H_1: \mu \neq 200, \delta \neq 50.$$

Расположим данные в порядке возрастания и вычислим наблюдаемые и ожидаемые кумулятивные частости (табл. 8.6).

Наблюдаемая кумулятивная частость, например 0,6000, показывает, что в трех из пяти случаев дневная заработная плата не превышает 249 руб. Поскольку генеральная совокупность специфицирована как нормальная с $\mu = 200$ и $\delta = 50$,

Таблица 8.6

Расчет критерия Колмогорова—Смирнова

Наблюдаемые значения	Частости	Кумулятивные частости		$\lvert F_i - E_i \rvert$
		наблюдаемые, F_i	ожидаемые, E_i	
146	0,2000	0,2000	0,1401	0,0599
231	0,2000	0,4000	0,7324	0,3324
249	0,2000	0,6000	0,8365	0,2365
288	0,2000	0,8000	0,9608	0,1608
291	0,2000	1,0000	0,9656	0,0344

ожидаемые частости находятся по таблице нормального распределения (табл. П.1 приложения). Например, для $x = 288$ находим $P\ (x \leqslant 288)$ следующим образом. Вычисляем: $Z = \frac{288 - 200}{50} = 1{,}76$. Затем по таблице стандартного нормального распределения (слева) находим площадь для $Z = 1{,}76$, которая составляет: 0,5000 + 0,4608 = 0,9608 площади нормальной кривой.

Максимальная разность $|F_i - E_i| = D = 0{,}3324$. По таблице критических значений D (табл. П.9 приложения) находим: $D_{крит} = 0{,}510$ при $\alpha = 0{,}10$ и $n = 5$. Поскольку $D_{факт} < D_{крит}$, нулевая гипотеза не может быть отклонена на 10%-ном уровне значимости.

Можно считать, что выборка работников проведена из нормально распределенной совокупности со средней величиной среднедневного заработка 200 руб./день и стандартным отклонением 50 руб./день.

Выбор закона распределения проводится на основе теоретического анализа. Кроме того, целесообразно руководствоваться следующей рекомендацией: выражение, определяющее функцию плотности распределения, должно зависеть от возможно меньшего числа параметров. Например, экспоненциальное распределение зависит от одного параметра — средней величины; нормальное и логнормальное распределение — от двух параметров.

8.3. Проверка гипотезы о связи на основе критерия χ^2 (хи-квадрат)

Одним из основных приложений критерия χ^2 является его использование при анализе таблиц сопряженности двух переменных для установления факта наличия и уровня значимости взаимосвязи. Как правило, критерий χ^2 применяется для анализа таблиц сопряженности номинальных признаков, однако он может быть использован и при анализе взаимосвязи порядковых или интервальных (количественных) переменных, несмотря на то, что для последних случаев существуют более мощные тесты.

Рассмотрим общий случай — таблицу сопряженности двух переменных размера $r \times s$. Обозначим:

n_{ij} — наблюдаемая частота (число объектов) в ячейке (i, j) таблицы, так называемая фактическая клеточная частота; $\hat{n}_{ij}$ — теоретически ожидаемая (по H_0) частота в этой ячейке, $i = 1, 2, ..., r, j - 1, 2, ..., s$; r — число строк; s — число столбцов.

$$\left.\begin{aligned} n_{i.} &= \sum_{j=1}^{s} n_{ij} \text{ — сумма по } i\text{-й строке} \\ n_{.j} &= \sum_{i=1}^{r} n_{ij} \text{ — сумма по } j\text{-му столбцу} \end{aligned}\right\} \text{маргинальные частоты;} \qquad (8.7)$$

$$n = \sum_{j=1}^{r} n_{.j} = \sum_{i=1}^{s} n_{i.} = \sum_{j=1}^{r}\sum_{i=1}^{s} n_{ij} \text{ — общее число объектов или объем выборки.}$$

Знак «.» означает суммирование по второму подстрочному знаку: по i или по j. В этом случае испытуемая гипотеза $H_0: n_{ij} = \hat{n}_{ij}$ или $H_0: \chi^2 = 0$, альтернативная гипотеза $H_1: n_{ij} \neq \hat{n}_{ij}$. Критерий χ^2 для проверки H_0 имеет вид:

$$\chi^2 = \sum_{(i)}\sum_{(j)} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}. \qquad (8.8)$$

Расчет теоретически ожидаемых частот в ячейках таблицы сопряженности должен проводиться, как мы уже указывали

выше, в предположении справедливости нулевой гипотезы. Нуль-гипотеза (H_0) в данном случае есть *предположение о статистической независимости* рассматриваемых переменных. Как известно из теории вероятностей, две случайные величины (события) являются статистически независимыми, если вероятность их совместной реализации равна произведению вероятностей реализации каждой из них по отдельности, т.е.

$$\pi(x_i, x_j) = \pi(x_i)\pi(x_j);$$

где

$$\pi(x_i, x_j) = \pi(\mathbf{X}_i = x_i, \mathbf{X}_j = x_j).$$

В нашем случае выборочными оценками соответствующих вероятностей π будут являться величины

$$p(x_i, x_j) = p(x_i) \cdot p(x_j),$$

$$p(x_i) = \frac{n_i}{n}, \; p(x_j) = \frac{n_j}{n},$$

и поэтому расчет теоретически ожидаемой частоты $\hat{n}_{ij}$ по H_0 следует проводить по формуле

$$\hat{n}_{ij} = n\frac{n_i}{n} \cdot \frac{n_j}{n} = \frac{n_i n_j}{n}, \qquad (8.9)$$

т.е. теоретическая частота равна произведению итогов по столбцу и строке, деленному на общий объем данных.

Если подставить выражение $\hat{n}_{ij}$ в формулу (8.9), то получим:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_i n_j} - 1\right). \qquad (8.10)$$

Используя формулу (8.10), мы можем находить эмпирические значения критерия χ^2 без промежуточного вычисления теоретических частот в явном виде.

Ясно, что для определения эмпирического значения критерия χ^2 нет необходимости рассчитывать все s теоретических частот в каждой строке, а достаточно найти лишь $s - 1$ значение частоты в $r - 1$ строке, так как оставшиеся частоты могут быть получены как разности между маргинальными суммами эмпирических частот и суммами известных теоретических частот, т.е. значения теоретических частот в последних строке и столбце таблицы всегда полностью детерминированы. Поэтому число степеней свободы для таблицы сопряженности $r \times s$ равно:

$$d.f. = (r - 1)(s - 1). \qquad (8.11)$$

Заметим, что для таблицы 2×2 число степеней свободы равно единице.

В таблице распределения статистики $\chi^2_{d.f.,\,\alpha}$ приведены значения этой величины для различных уровней значимости при различных числах степеней свободы (табл. П.4 приложения). Например, на уровне $\alpha = 0{,}001$ для $d.f. = 1$ мы находим: $\chi^2 = 10{,}827$. Это означает, что равное или большее значение величины χ^2 может встретиться только один раз из 1000 при условии, что все сделанные допущения (нуль-гипотеза) справедливы. Другими словами, если выполняется предположение об отсутствии взаимосвязи между переменными, то крайне маловероятно ($P < 0{,}001$), что наблюдаемые и ожидаемые частоты будут отличаться настолько, что фактическая величина χ^2 будет равной или большей 10,827. Если же $\chi^2_{\text{факт}} \geqslant \chi^2_{d.f.,\,\alpha}$, то гипотеза H_0 на данном уровне значимости α может быть отвергнута.

Вероятность того, что, отвергая нулевую гипотезу, мы совершаем ошибку (первого рода), которая численно равна уровню значимости α, задаваемому при проверке гипотезы, обычно принимается равной 0,05 или 0,01, или 0,001.

Интерпретация χ^2 теста зачастую усложняется, когда в таблице сопряженности имеются ячейки с нулевыми значениями наблюдаемых частот. Дело в том, что если пара (x_i, x_j) значений переменных не наблюдалась в выборке, то это может означать, что объем выборки не столь велик, чтобы зафиксировать такую редкую комбинацию, либо что данная комбинация невозможна по каким-то объективным причинам. В последнем случае действительное число степеней свободы анализируемой системы меньше числа степеней свободы таблицы сопряженности, на основе которого проведена оценка уровня значимости χ^2 теста.

Корректировка применения χ^2 теста возможна лишь в том случае, если эмпирические данные в таблице сопряженности

есть результаты независимой случайной выборки относительно большого объема n. Последнее требование вызвано тем, что выборочное распределение χ^2 аппроксимирует табличное распределение статистики χ^2 только при больших n. Естественно, возникает вопрос о том, насколько велико должно быть n, чтобы иметь возможность использовать данный тест. Ответ на этот вопрос зависит от числа ячеек и величин маргинальных сумм. Вообще говоря, чем меньше число ячеек и чем более близки между собой по величине маргиналы, тем меньше может быть n. Существует, однако, критическое число, позволяющее оценить снизу по n диапазон вероятностного применения критерия χ^2: если в данной таблице сопряженности любая из теоретических ожидаемых частот $\hat{n}_{ij}$ в ячейке (i, j) не больше пяти, то рекомендуется провести, если это возможно, модификацию таблицы либо воспользоваться другим критерием.

В общем случае корректировка таблицы размера $r \times s$ затруднительна. Практика показала, что если число ячеек велико, а ожидаемые частоты, равные или меньше пяти, встречаются лишь в одной-двух ячейках, то проведение корректировки нецелесообразно; в остальных случаях разумной альтернативой является объединение категорий (градаций) с тем, чтобы элиминировать подобные ячейки. Разумеется, подобное объединение должно быть таким, чтобы полученная в результате комбинация не была содержательно бессмысленной.

Пример. Согласно опросу 157 предпринимателей, работающих в приватизированных кафе и ресторанах, об оценке возможностей деятельности при разных формах собственности получены следующие данные (табл. 8.7).

Испытаем гипотезу о независимости переменных H_0: $n_{ij} = \hat{n}_{ij}$, где $\hat{n}_{ij}$ — генеральные частоты, оценками которых выступают выборочные частоты n_{ij}. Теоретические частоты, рассчитанные в соответствии с нуль-гипотезой, как $\hat{n}_{ij} = \frac{n_i n_j}{n}$, представлены в табл. 8.8.

Таким было бы распределение ответов о возможностях деятельности, если бы формы собственности никак не сказывались. Задавая уровень значимости $\alpha = 0{,}05$, находим по

Таблица 8.7

Исходные данные: таблица сопряженности

Организационно-правовая форма	Оценка возможностей деятельности					
	крайне неблагоприятно	неблагоприятно	трудно сказать	благоприятно	исключительно благоприятно	Итого
Один владелец	18	16	5	13	5	57
Товарищество	4	4	10	11	11	40
Товарищество с ограниченной ответственностью	10	15	8	23	4	60
Итого	32	35	23	47	20	157

табл. П.4 приложения критическое значение критерия $\chi^2_{d.f.,\,\alpha}$ при числе степеней свободы $d.f. = (3-1)(5-1) = 8$. Отсюда: $\chi^2_{d.f.,\,\alpha} = 15{,}51$.

Различия между фактическими и теоретическими клеточными частотами обобщаются в величине χ^2:

$$\chi^2 = \frac{(18-11{,}6)^2}{11{,}6} + \frac{(16-12{,}7)^2}{12{,}7} + \frac{(5-8{,}3)^2}{8{,}3} + \frac{(13-17{,}1)^2}{17{,}1} +$$

$$+ \frac{(5-7{,}3)^2}{7{,}3} + \frac{(4-8{,}2)^2}{8{,}2} + \ldots + \frac{(23-18)^2}{18} + \frac{(4-7{,}6)^2}{7{,}6} = 31{,}045.$$

Таблица 8.8

Теоретические частоты

Организационно-правовая форма	Оценка возможностей деятельности					
	крайне неблагоприятно	неблагоприятно	трудно сказать	благоприятно	исключительно благоприятно	Итого
Один владелец	11,6	12,7	8,3	17,1	7,3	57
Товарищество	8,2	8,9	5,9	11,9	5,1	40
Товарищество с ограниченной ответственностью	12,2	13,4	8,8	18,0	7,6	60
Итого	32	35	23	47	20	157

Так как $\chi^2_{факт} > \chi^2_{крит}$, H_0 отклоняется, т.е. форма собственности небезразлична для деятельности кафе и ресторанов. Таким образом, наблюдаемое значение χ^2 является *значимым* на 5%-ном уровне значимости, и нулевая гипотеза может быть отвергнута в пользу альтернативной.

Для таблиц 2 × 2 величина χ^2 вычисляется с учетом поправки Ф. Йейтса:

$$\chi^2 = \sum_{(i)}\sum_{(j)} \frac{(|n_{ij} - \hat{n}_{ij}| - 0{,}5)^2}{\hat{n}_{ij}}. \qquad (8.12)$$

Вычитание 0,5 из каждой величины $|n_{ij} - \hat{n}_{ij}|$ называется поправкой на непрерывность, так как оно корректирует несоответствие между дискретным биномиальным распределением при числе степеней свободы, равном единице, и непрерывным распределением χ^2.

Итак, мы рассмотрели один из возможных способов ответа на вопрос: существует ли связь между двумя переменными? Для этого нам понадобилось выдвинуть нулевую гипотезу, что такой связи нет, а затем рассмотреть способ статистического испытания этой гипотезы. Можно оценить величину риска в принятии предположения о *существовании* связи. Но означает ли это, что данная связь существенна с точки зрения ее силы? Вовсе не обязательно. Вопрос о силе или степени, тесноте зависимости — это иной вопрос, отличный от вопроса о существовании взаимосвязи.

В социально-экономических исследованиях, как правило, установление факта наличия связи между переменными не самоцель. Установив наличие связи, исследователь должен измерить ее силу (тесноту) с тем, чтобы иметь возможность сравнивать взаимосвязи между различными характеристиками, выделять наиболее сильные из них (гл. 9, 11).

8.4. Проверка гипотезы о средних величинах

Основные гипотезы о средних величинах следующие: гипотезы о значении генеральной средней (при известной генеральной дисперсии или при неизвестной генеральной дисперсии); гипотезы о равенстве генеральных средних нормально распределенных совокупностей (при известных генеральных дисперсиях, при неизвестных равных генеральных дисперсиях, при неизвестных неравных генеральных дисперсиях).

Первая задача чаще всего решается при неизвестной генеральной дисперсии. Испытуемая гипотеза H_0: $\mu = \mu_0$, альтернативная гипотеза H_1: $\mu \neq \mu_0$. Испытание гипотезы проводят с помощью t-критерия. При большом числе наблюдений критическое значение критерия определяется по таблице интеграла вероятностей (табл. П.1 приложения), при малом — по таблице распределения Стьюдента с заданным уровнем значимости и числом степеней свободы $n - 1$.

Если испытуемая гипотеза H_0: $\mu = a$, то фактическое значение критерия представляет отношение оцениваемой разности к средней возможной ошибке выборочной средней

$$t = \frac{\bar{x} - a}{s_{\bar{x}}}, \qquad (8.13)$$

где $s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$ — при большой выборке;

$s_{\bar{x}} = \sqrt{\frac{s^2}{n-1}}$ — при малой выборке.

Если $t_{факт} < t_{крит}$, H_0 не отклоняется; если $t_{факт} > t_{крит}$, H_0 отклоняется.

Пример. Часовая выработка забойщика при добыче угля в шахте по норме составляет 400 кг. Фактическая выработка соответствовала норме. При переходе в новый забой условия работы забойщиков усложнились. Для проверки обоснованности нормы в новых условиях был проведен учет работы девяти забойщиков: их средняя часовая выработка составила 388 кг с дисперсией, равной: $s^2 = 171$.

Выдвигается гипотеза, что норму выработки пересматривать не нужно, т.е. H_0: $\mu = 400$ кг. Проверим эту гипотезу на 5%-ном уровне значимости. Критическое значение t-критерия определяется по таблице распределения Стьюдента при доверительной вероятности 0,95 (1 − 0,05) и числе степеней свободы $d.f. = n - 1 = 8$. Критическое значение составит: $t_{крит} = 2{,}3$. Фактическое значение t-критерия вычисляется по формуле (8.13):

$$t = \frac{388 - 400}{\sqrt{\frac{171}{8}}} = 2{,}6.$$

Поскольку $t_{факт} > t_{крит}$, H_0 отклоняется. Норма выработки в новых условиях должна быть пересмотрена, так как производительность труда стала существенно ниже нормативной.

В рассмотренном примере различие между фактическим и табличным значениями t-критерия невелико, поэтому вывод недостаточно надежен. Надежность вывода вообще понижается, если нет уверенности в нормальном распределении генеральной совокупности.

Гипотеза о равенстве средних может рассматриваться как гипотеза о связи, если сопоставляются средние величины, обусловленные действием какого-либо фактора. Например, сравнивается средняя заработная плата инженеров двух специальностей. Нулевая гипотеза состоит в том, что специальность рабочего не влияет на заработок. Если выяснится, что $t_{факт} > t_{крит}$, нулевую гипотезу отклоняют и делают вывод, что специальность оказывает влияние на заработную плату.

Рассмотрим решение этой задачи при условии, что генеральные дисперсии неизвестны, но принимаются равными. При сравнении средних величин выдвигается гипотеза, что обе выборки принадлежат одной и той же генеральной совокупности со средней μ и дисперсией σ^2.

При неизвестной генеральной дисперсии формула t-критерия имеет вид:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \qquad (8.14)$$

Поскольку s_1^2 и s_2^2 рассматриваются как выборочные оценки общей дисперсии σ^2, то формула (8.14) может быть записана так:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(n_1^{-1} + n_2^{-1})}}, \qquad (8.15)$$

где $\bar{x}_1$, $\bar{x}_2$ — выборочные средние;

s^2 — выборочная оценка общей дисперсии;

$$s^2 = \frac{\Sigma(x_{i1} - \bar{x}_1)^2 + \Sigma(x_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2}. \qquad (8.16)$$

Гипотеза H_0 отклоняется, если $|t| > t_{1-\alpha,\, d.f. = n_1 + n_2 - 2}$.

Пример. Для проверки устойчивости цен на яблоки в летний период на двух рынках города проведено выборочное обследование: на первом рынке по данным 15 продавцов определена средняя цена, равная 20 руб./кг, при среднем квадратическом отклонении $s_1 = 5$ руб.; на втором рынке обследованы 17 продавцов, средняя цена оказалась равной 25 руб./кг, $s_2 = 4$ руб.

$$H_0: \mu_1 = \mu_2,\ H_1: \mu_1 \neq \mu_2;$$

$$t = \frac{|20 - 25|}{\sqrt{\frac{375 + 272}{15 + 17 - 2}}} \cdot \sqrt{\frac{15 \cdot 17}{15 + 17}} = 3{,}04.$$

При $\alpha = 0{,}05$ и $d.f. = 30$, $t_{крит} = 2{,}042$. Так как $t_{факт} > t_{крит}$, то H_0 отклоняется, т.е. различия в ценах на двух рынках нельзя объяснить лишь случайностями выборки.

Проверка той же нулевой гипотезы при односторонней критической области будет проводиться на следующих условиях определения: $t_{крит} = 1 - 2\alpha$ и $d.f. = n_1 + n_2 - 2$. Следовательно, если H_1: $\mu_1 = \mu_2$, $t_{крит} = 1{,}697 (2\alpha = 0{,}1,\ d.f. = 30)$, так что H_0 опять-таки отклоняется.

Проверка гипотезы о средних величинах при неизвестных дисперсиях, равенство которых не предполагается, здесь не рассматривается ввиду недостаточной теоретической разработанности[1].

8.5. Основы дисперсионного анализа

Если ставится задача сравнения двух выборочных дисперсий, то для ее решения применяется критерий, названный в честь английского статистика Р. Фишера (1890—1969) F-критерием. Этот критерий представляет собой отношение выборочных дисперсий s_1^2 и s_2^2, которые рассматриваются как оценки одной и той же генеральной дисперсии σ^2:

$$F = \frac{s_1^2}{s_2^2}. \qquad (8.17)$$

[1] *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ: Пер. с англ. / Под ред. Г. П. Башарина. — М.: Мир, 1982.

Испытуемая гипотеза является нулевой гипотезой H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$, альтернативная гипотеза H_1: $\sigma_1^2 \neq \sigma_2^2 \neq \sigma^2$.

F-критерий строится так, что в числителе стоит бóльшая дисперсия. $F_{\min} = 1$, $F_{\max} \to \infty$. Критические значения *F*-критерия берутся из таблиц *F*-распределения. *F*-распределение зависит от уровня значимости и от числа степеней свободы сравниваемых дисперсий $d.f._1$ и $d.f._2$ (табл. П.3 приложения).

В дисперсионном анализе общая вариация подразделяется на составляющие и проводится сравнение этих составляющих. Испытуемая гипотеза заключается в том, что если данные каждой группы представляют случайную выборку из нормально распределенной генеральной совокупности, то величины всех частных дисперсий должны быть пропорциональны своим степеням свободы и каждую из них можно рассматривать как оценку генеральной дисперсии.

Дисперсионный анализ часто применяется совместно с аналитической группировкой (см. гл. 6). В этом случае данные подразделяются на группы по значениям признака-фактора, вычисляются значения средних величин результативного признака в группах, считается, что различия в их значениях определяются различиями в значениях фактора. Задача состоит в оценке существенности различий между средними значениями результативного признака в группах. Итак, испытуемая гипотеза может быть записана как гипотеза о средних величинах H_0: $\mu_1 = \mu_2 = \mu_k = \ldots$ Как было показано в подразд. 8.4, когда выделяются две группы, эта задача решается с помощью *t*-критерия. Если же число сравниваемых групп больше двух, то существенность различий между группами доказывается с помощью дисперсионного анализа, на основе *F*-критерия. Заметим, что результаты дисперсионного анализа, так же как и выводы о характере связи, значения показателей ее силы и тесноты, зависят от числа групп, выделенных по признаку-фактору.

В случае выделения групп по одному фактору мы имеем так называемый *однофакторный дисперсионный комплекс*. Разложение дисперсии при этом проводится в соответствии с правилом сложения дисперсий (см. гл. 6):

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2 = \sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2 n_j + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2,$$

где y_{ij} — значение результативного признака у *i*-й единицы в *j*-й группе;
i — номер единицы, $i = 1, \ldots, n_j$;
j — номер группы;
n_j — численность *j*-й группы;
$\bar{y}_j$ — средняя величина результативного признака в *j*-й группе;
$\bar{y}$ — общая средняя результативного признака.

Если обозначить суммы квадратов отклонений буквой *D*, получим равенство:

$$D_{\text{общ}} = D_{\text{факт}} + D_{\text{ост}}. \qquad (8.18)$$

На основе разложения дисперсии (8.18) в соответствии с гипотезой отсутствия различий между группами могут быть получены три оценки генеральной дисперсии, пропорциональные степени свободы: на основе общей вариации, межгрупповой (факторной) и внутригрупповой (остаточной). Число степеней свободы равно:

для общей вариации $d.f._{\text{общ}} = (mn_j - 1) = n - 1$;

для межгрупповой вариации $d.f._{\text{факт}} = m - 1$;

для внутригрупповой вариации $d.f._{\text{ост}} = (n_j - 1)m = n - m$.

Как и суммы квадратов отклонений, числа степеней свободы связаны между собой равенством

$$d.f._{\text{общ}} = d.f._{\text{факт}} + d.f._{\text{ост}}$$

или

$$n - 1 = (m - 1) + (n - m). \qquad (8.19)$$

Деление сумм квадратов отклонений на соответствующее число степеней свободы дает три оценки генеральной дисперсии σ^2:

$$s^2_{\text{общ}} = \frac{D_{\text{общ}}}{n-1}; \qquad (8.20)$$

$$s^2_{\text{факт}} = \frac{D_{\text{факт}}}{m-1};$$

$$s^2_{\text{ост}} = \frac{D_{\text{ост}}}{n-m}.$$

Поскольку $D_{\text{факт}}$ измеряет вариацию результативного признака, связанную с изменением фактора, по которому проведена группировка, а $D_{\text{ост}}$ — вариацию, связанную с изменением всех прочих факторов, сравнение этих величин, рассчитанных на одну степень свободы, дает возможность оценить существенность влияния признака-фактора на результативный признак с помощью F-критерия:

$$F = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}}.$$

Эта запись предполагает, что $s^2_{\text{факт}} \geqslant s^2_{\text{ост}}$. Как правило, мы получаем именно такое соотношение. Если $F_{\text{факт}} > F_{\text{табл}}(\alpha, d.f.1, d.f.2)$, можно утверждать, что нуль-гипотеза не соответствует фактическим данным, влияние признака-фактора является существенным или, иначе говоря, статистически значимым.

Этапы однофакторного дисперсионного анализа представлены в табл. 8.9.

По данным табл. 8.10 проверим гипотезу $H_0 : \mu_1 = \mu_2 = \mu_3$, т.е. предположим, что оборачиваемость средств никак не влияет на прибыль.

Таблица 8.9

Схема однофакторного дисперсионного анализа

Источник вариации	Сумма квадратов отклонений, D	Число степеней свободы, $d.f.$	Средний квадрат отклонений, $s^2 = D : d.f.$	F-критерий
Между группами	$\sum_{1}^{m}(\bar{y}_j - \bar{\bar{y}})^2 n_j$	$m - 1$	s^2_1	—
Внутри групп	$\sum_{1}^{m}\sum_{1}^{n_j}(y_{ij} - \bar{y}_j)^2$	$n - m$	s^2_2	$F = \frac{s^2_1}{s^2_2}$
Общая	$\sum_{1}^{m}\sum_{1}^{n_j}(y_{ij} - \bar{\bar{y}})^2$	$n - 1$	s^2	—

Таблица 8.10

Однофакторный дисперсионный анализ: зависимость прибыли от средней продолжительности оборота оборотных средств

Средняя продолжительность оборота, дней	Число предприятий	Прибыль за год, млн руб.	Среднегодовая прибыль, млн руб.
x_i	n_j	y_{ij}	y_j
11—30	6	10,0; 9,5; 14,8; 20,3; 18,7; 14,3;	14,60
31—50	8	19,3; 12,1; 20,5; 13,5; 11,5; 8,3; 9,8; 8,6	12,95
51—70	6	9,0; 9,5; 7,3; 8,5; 5,2; 4,9	7,40
Итого	20		11,78

По данным табл. 8.10 находим:

$D_{\text{факт}} = 173{,}7722$ $\quad d.f._{\text{факт}} = 3 - 1 = 2$ $\quad s^2_{\text{факт}} = 86{,}886$

$D_{\text{ост}} = 266{,}8$ $\quad d.f._{\text{ост}} = 20 - 3 = 17$ $\quad s^2_{\text{ост}} = 15{,}694$

$D_{\text{общ}} = 440{,}572$ $\quad d.f._{\text{общ}} = 20 - 1 = 19$

Отсюда фактическое значение F-критерия составляет:

$$F = 86{,}886 : 15{,}694 = 5{,}54.$$

Критическое значение F-критерия из табл. П.3 приложения равно:

$$F_{\text{крит}}(\alpha = 0{,}05,\ d.f._1 = 2,\ d.f._2 = 17) = 3{,}59.$$

Поскольку $F_{\text{факт}} > F_{\text{табл}}$, H_0 отклоняется.

На 5%-ном уровне значимости нельзя признать справедливой гипотезу о равенстве годовой прибыли по группам предприятий с разной продолжительностью оборота оборотных средств. Скорость оборота средств является важным фактором формирования прибыли, на это указывало и ранее вычисленное значение эмпирического корреляционного отношения: $\eta = 0{,}628$.

Рассмотрим *двухфакторный дисперсионный анализ*, основой проведения которого служит комбинационная группировка по двум факторам x и z, с последующим разложением дисперсии результативного признака y:

$$\sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y})^2 = \sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2 n_j + \sum_{k=1}^{p}(\bar{y}_k-\bar{y})^2 n_k +$$

$$+\sum_{j=1}^{m}\sum_{k=1}^{p}(\bar{y}_{jk}-\bar{y}_j-\bar{y}_k+\bar{y})^2 n_{jk} + \sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{ik})^2, \quad (8.21)$$

где n_{jk} — число единиц в группе, образованной комбинацией j-го значения признака x и k-го значения признака z;

n — общее число единиц, $n = \sum_{j=1}^{m} n_j = \sum_{k=1}^{p} n_k = \sum_{j=1}^{m}\sum_{k=1}^{p} n_{jk}$;

i — номер единицы в j-й группе по признаку x и в k-й по признаку z;

$j = \overline{1, m}$; $k = \overline{1, p}$;

$\bar{y}_{jk}$ — среднее значение признака y в группе, образованной комбинацией j-го значения признака x и k-го значения признака z;

$\bar{y}_j$ — среднее значение признака y в j-й группе по признаку x;

$\bar{y}_k$ — среднее значение признака y в k-й группе по признаку z;

$\bar{y}$ — общая средняя признака y в целом по выборке;

n_j — число единиц в j-й группе по признаку x;

n_i — число единиц в k-й группе по признаку z.

Равенство (8.21) можно записать так:

$$D_{общ} = D_x + D_z + D_{xz} + D_{ост}, \quad (8.22)$$

где D_x — вариация y под влиянием фактора x;

D_z — вариация y под влиянием фактора z;

D_{xz} — вариация y, обусловленная взаимодействием факторов x и z;

$D_{ост}$ — вариация y под влиянием прочих факторов.

Первые три слагаемых составляют вариацию признака y, вызванную изучаемыми факторами, поэтому равенство (8.22) можно записать в виде:

$$D_{общ} = D_{факт} + D_{ост}, \quad (8.23)$$

$$D_{факт} = D_x + D_z + D_{xz}. \quad (8.24)$$

Величина $D_{факт}$ может быть рассчитана не через составляющие, а непосредственно как

$$D_{факт} = \sum_{j=1}^{m}\sum_{k=1}^{p}(\bar{y}_{jk}-\bar{y})^2 n_{jk}. \quad (8.25)$$

Однако при неравенстве численностей подгрупп n_{jk} и групп n_j и n_k равенство (8.25) нарушается (за счет взвешивания при неравных весах).

Поэтому рассчитываются невзвешенные величины:

$$D'_{факт} = \sum_{j=1}^{m}\sum_{k=1}^{p}(\bar{y}_{jk}-\bar{y})^2; \quad (8.26)$$

$$D'_x = \sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2;$$

$$D'_z = \sum_{k=1}^{p}(\bar{y}_k-\bar{y})^2;$$

$$D'_{xz} = D'_{факт} - D'_x - D'_z.$$

Затем на основе сравнения взвешенной (8.25) и невзвешенной величин факторной дисперсии находят поправочный коэффициент:

$$K = D_{факт} : D'_{факт}. \quad (8.27)$$

Этот коэффициент используется для корректировки невзвешенных сумм квадратов отклонений D'_x, D'_z, D'_{xz}, на основе которых проводят расчет F-критериев:

$$D_{x(корр)} = D'_x K;\ D_{z(корр)} = D'_z K;\ D_{xz(корр)} = D'_{xz} K.$$

Число степеней свободы для каждой суммы квадратов отклонений составляет:

$$d.f._x = m - 1;\ d.f._z = p - 1;$$
$$d.f._{xz} = (m-1)(p-1) = mp - m - p + 1;$$

в целом:

$$d.f._{факт} = d.f._x + d.f._z + d.f._{xz} = mp - 1;$$
$$d.f._{общ} = n - 1;\ d.f._{ост} = d.f._{общ} - d.f._{факт} = n - mp. \quad (8.28)$$

В двухфакторном дисперсионном анализе испытуемые гипотезы формулируются следующим образом:

1) H_0: $\mu_{1.} = \mu_{2.} = \ldots = \mu_{m.}$;

2) H_0: $\mu_{.1} = \mu_{.2} = \ldots = \mu_{.p}$;

3) H_0: $\mu_{11} = \mu_{12} = \mu_{22} = \ldots = \mu_{mp}$.

Вся процедура двухфакторного дисперсионного анализа обобщается в табл. 8.11.

Таблица 8.11

Схема двухфакторного дисперсионного анализа

Источник вариации	Сумма квадратов отклонений, D	Число степеней свободы, $d.f.$	Средний квадрат отклонений, $s^2 = D/d.f.$	F-критерий
Факторы x и z	$D'_{\text{факт}} \cdot K$	$mp - 1$	$s^2_{\text{факт}}$	$F = \frac{s_x^2}{s^2_{\text{ост}}}$
Фактор x	$D'_x \cdot K$	$m - 1$	s_x^2	
Фактор z	$D'_z \cdot K$	$p - 1$	s_z^2	$F = \frac{s_z^2}{s^2_{\text{ост}}}$
Взаимодействие факторов x и z	$(D'_{\text{факт}} - D'_x - D'_z)K$	$mp - p - m + 1$	s_{xz}^2	$F = \frac{s_{xz}^2}{s^2_{\text{ост}}}$
Остаточная	$D_{\text{общ}} - D'_{\text{факт}} \cdot K$	$n - mp$	$s^2_{\text{ост}}$	—
Общая	$D_{\text{общ}}$	$n - 1$	s^2	—

Решение о первой гипотезе принимается на основе сравнения

$F = s_x^2/s^2_{\text{ост}}$ с $F_{\text{крит}(\alpha,\, d.f._1 = d.f._x;\; d.f._2 = d.f._{\text{ост}})}$. Если $F_{\text{факт}} > F_{\text{крит}}$, то H_0 отклоняется.

Вторая гипотеза испытывается на основе сравнения

$$F = s_z^2/s^2_{\text{ост}} \text{ с } F_{\text{крит}(\alpha,\, d.f._1 = d.f._x;\; d.f._2 = d.f._{\text{ост}})}.$$

Третья — на основе сравнения

$$F = s_{xz}^2/s^2_{\text{ост}} \text{ с } F_{\text{крит}(\alpha,\, d.f._1 = d.f._{xz};\; d.f._2 = d.f._{\text{ост}})}.$$

Во всех случаях, если $F_{\text{факт}} > F_{\text{крит}}$, H_0 отклоняется.

На основе F-критерия принимаются решения о форме уравнения регрессии, о статистической значимости той или иной объясняющей переменной при построении многофакторного уравнения регрессии (гл. 9) и др.

Пример. Проанализируем зависимость прибыли от продолжительности оборота и величины запаса оборотных

Таблица 8.12

Исходные данные: двухфакторный дисперсионный анализ

Среднегодовой запас оборотных средств, млн руб.	Средняя продолжительность оборота, дней	Число предприятий	Прибыль за год, млн руб.
z_k	x_j	n_{jk}	y_{ijk}
55—85	11—30	1	14,3
	31—50	2	9,8; 11,5
	51—70	1	7,3
85—115	11—30	2	9,5; 14,8
	31—50	4	12,1; 8,3; 8,6; 19,3
	51—70	2	8,5; 5,2
115—145	11—30	3	20,3; 18,7; 10
	31—50	2	20,5; 13,5
	51—70	3	9,0; 9,5; 4,9

средств по данным 20 предприятий, используя метод двухфакторного дисперсионного анализа. Исходные данные представлены в табл. 8.12.

Среднегодовая прибыль по группам и подгруппам представлена в табл. 6.6 и 6.7 (см. гл. 6).

Рассчитаем колеблемость прибыли с изменением запаса оборотных средств (табл. 8.13).

Вычислим невзвешенные суммы квадратов отклонений D'_x (по данным табл. 6.6) и D'_z (по данным табл. 8.13):

Таблица 8.13

Зависимость прибыли от величины среднегодового запаса оборотных средств

Среднегодовой запас оборотных средств, млн руб.	Число предприятий	Среднегодовая прибыль за год, млн руб.	Расчет межгрупповой дисперсии
z_k	n_k	$\bar{y}_k$	$(\bar{y}_k - y)^2 n_k$
55—85	4	10,7250	4,4521
85—115	8	10,7875	7,8805
115—145	8	13,3000	18,4832
Итого	20	11,7800	30,8158

$$D_x' = 28{,}5057;\ D_z' = 4{,}4085.$$

По данным табл. 6.7 рассчитаем взвешенное и невзвешенное значение факторной колеблемости:

$$D_{\text{факт}} = 242{,}3537 \text{ (взвешенная)};$$
$$D_{\text{факт}}' = 115{,}9851 \text{ (невзвешенная)}.$$

Отсюда поправочный коэффициент равен:

$$K = D_{\text{факт}} : D_{\text{факт}}' = 242{,}351 : 115{,}9851 = 2{,}0894.$$

Находим невзвешенную величину колеблемости y за счет взаимодействия факторов x и z:

$$D_{xz}' = D_{\text{факт}}' - D_x' - D_z' = 115{,}9851 - 28{,}5057 - 4{,}4085 = 83{,}0709.$$

Вычислим скорректированные значения составляющих $D_{\text{факт}}$:

$$D_{x\,(\text{скорр})} = D_x' \cdot K = 28{,}5057 \cdot 2{,}0895 = 59{,}5627;$$
$$D_{z(\text{скорр})} = D_z' \cdot K = 4{,}4085 \cdot 2{,}0895 = 9{,}2116;$$
$$D_{xz(\text{скорр})} = D_{xz}' \cdot K = 83{,}0709 \cdot 2{,}0895 = 173{,}5767.$$

Суммируем полученные величины, т.е. вычисляем факторную колеблемость:

$$D_{\text{факт}} = 59{,}5627 + 9{,}2116 + 173{,}5767 = 242{,}351.$$

Это значение практически не отличается от того, которое было получено по данным табл. 6.7 ($D_{\text{факт}}' = 242{,}3537$).

Теперь мы имеем все необходимое для таблицы двухфакторного дисперсионного анализа в соответствии со схемой, приведенной в табл. 8.14.

Все фактические значения F-критерия меньше табличных (табл. П.3 приложения): $1{,}68 < 2{,}95$; $1{,}65 < 3{,}98$; $2{,}41 < 3{,}36$. Так что, строго говоря, влияние изучаемых факторов на прибыль по данной выборке не доказано. «Чистое» воздействие фактора продолжительности оборота средств оказалось слабым, поскольку большее значение имеет взаимодействие выделенных факторов, которое в однофакторном анализе присоединилось к оценке воздействия оборачиваемости. Все нулевые гипотезы не могут быть отклонены на 5%-ном уровне значимости.

Таблица 8.14

Пример двухфакторного дисперсионного анализа

Источник вариации	Сумма квадратов отклонений, D	Число степеней свободы, $d.f.$	Средний квадрат отклонений, $s^2 = D/d.f.$	F-критерий
Факторы x и z	$D_{\text{факт}} = 242{,}351$	$d.f._{\text{факт}} = 9 - 1 = 8$	$s^2_{\text{факт}} = 30{,}29$	$F_{\text{факт}} = 1{,}68$
Фактор x	$D_{\text{скорр}} = 59{,}5627$	$d.f._x = 3 - 1 = 2$	$s_x^2 = 29{,}78$	$F_x = 1{,}65$
Фактор z	$D_{\text{скорр}} = 9{,}2116$	$d.f._z = 3 - 1 = 2$	$s_z^2 = 4{,}6058$	$F_z = \frac{4{,}60}{18{,}02}$
Взаимодействие факторов x и z	$D_{\text{скорр}} = 173{,}5767$	$d.f._{xz} = 9 - 3 - 3 + 1 = 4$	$s_{xz}^2 = 43{,}39$	$F_{xz} = 2{,}41$
Остаточная	$D_{\text{ост}} = 198{,}221$	$d.f._{\text{ост}} = 20 - 9 = 11$	$s^2_{\text{ост}} = 18{,}02$	—
Общая	$D_{\text{общ}} = 440{,}572$	$d.f._{\text{общ}} = 20 - 1 = 19$	$s^2_{\text{общ}} = 23{,}188$	—

Рассмотренные направления проверки статистических гипотез охватывают лишь важнейшие из них. Процедура испытания статистических гипотез применяется для определения того, случайно или нет полученное значение коэффициента корреляции, коэффициента вариации и т.д., случайны или нет различия в значениях показателей (медиан, коэффициентов корреляции, регрессии и т.д.) в разных совокупностях. Во всех случаях результатом является *вероятностное суждение*, которое составляет сущность анализа данных в разнообразных сферах: в медицине, биологии, технике, политике, спорте, экономике, психологии и социологии.

8.6. Некоторые непараметрические критерии

В предыдущих подразделах рассмотрено применение основных статистико-математических критериев: хи-квадрата (непараметрический критерий) и F-критерия (параметрический критерий). В этом подразделе рассмотрим дополнительно ряд непараметрических критериев, актуальность использования которых непрерывно возрастает.

Непараметрическое тестирование не нуждается в каких-либо предположениях относительно характера распределения генеральной совокупности, из которой взята изучаемая выборка. Это наиболее неприятный момент для параметрических тестов, которые выведены в предположении о нормальности генеральной совокупности. При сравнении двух и более генеральных совокупностей предполагается, что генеральные дисперсии равны. Большинство параметрических тестов требуют, чтобы данные были представлены в интервальной шкале или шкале отношений, в то время как многие непараметрические тесты не включают таких требований к данным.

Непараметрические тесты используются вместо параметрических, когда данные измерены на номинальной или порядковой шкале; когда данные измерены на интервальной или порядковой шкале, но предположение о нормальности не может быть сделано.

По сравнению с параметрическими тестами непараметрическое тестирование имеет следующие преимущества и недостатки.

Преимущества

1. Меньше предположений о генеральной совокупности. Наиболее важное из них то, что совокупность не должна быть нормально распределенной или приблизительно нормальной. Непараметрические тесты не включают никаких предположений о каком-либо типе распределения.

2. Методы непараметрического тестирования могут быть применены даже тогда, когда выборка очень мала.

3. Могут использоваться данные, представленные в любых шкалах измерения (номинальные, порядковые).

4. Простота вычислений, которые могут проводиться на микрокалькуляторе. Это прежде всего связано с малым числом наблюдений, к которым применяются непараметрические тесты.

Недостатки

1. По сравнению с параметрическими тестами информация, имеющаяся в данных, используется менее эффективно, и мощность тестов ниже, чем параметрических. По этой причине параметрические тесты предпочтительнее, когда требуемые предположения относительно генеральной совокупности могут быть сделаны.

2. Непараметрическое тестирование больше зависит от статистических таблиц, если не используется специальный пакет прикладных программ.

Критерий знаков Вилкоксона (случай одной выборки)

Для одной выборки критерий знаков Вилкоксона применяется, когда в отношении генеральной совокупности может быть выдвинута гипотеза о медиане. Никаких предположений о характере распределения не делается, кроме того, что совокупность примерно симметрична. Данные могут быть измерены на интервальной шкале или шкале отношений, или на порядковой шкале. В последнем случае критерий знаков Вилкоксона незаменим.

Рассмотрим методику применения этого критерия.

Нулевая и альтернативная гипотезы могут быть представлены в виде табл. 8.15.

Порядок расчета критерия Вилкоксона, W.

1. Для всех наблюдаемых величин рассчитываются разности $d_i = x_i - m_0$.

2. Исключаются наблюдения, для которых $d_i = 0$, остальные значения $|d_i|$ ранжируются так, что наименьшему значению присваивается ранг 1. В случае появления связанных рангов они рассчитываются как средние из соответствующей суммы мест и таким образом используются в расчете.

3. Для наблюдений, у которых $x_i > m_0$, ранги записываются в особую колонку R^+.

4. Рассчитывается критерий W как сумма значений колонки R^+, т. е. $W = \Sigma R^+$.

Таблица 8.15

Методика применения теста знаков Вилкоксона

Гипотеза	Двусторонний тест	Левосторонний тест	Правосторонний тест
Нулевая, H_0	$m = m_0$	$m \geqslant m_0$	$m \leqslant m_0$
Альтернативная, H_1	$m \neq m_0$	$m < m_0$	$m > m_0$

Примечание. Здесь m — медиана в генеральной совокупности; m_0 — значение, которое проверяется.

5. Для различных уровней значимости даются верхнее и нижнее значения на заданном n (где n — число наблюдений, для которых $d_i \neq 0$). Критические значения критерия знаков Вилкоксона приведены в табл. П.10 приложения. Область отклонения H_0 может быть либо с одной, либо с двух сторон в зависимости от того, какая нулевая гипотеза испытывается. В случае отсутствия специальных таблиц W-статистики может быть использовано стандартное нормальное распределение, т. е. Z-статистика с учетом n.

Пример. Представитель муниципалитета потребовал, чтобы питьевая вода содержала менее чем 40 частиц различных металлов на миллион, в соответствии с рекомендациями Комитета по здравоохранению. Для того чтобы проверить, отвечает ли питьевая вода предъявленным требованиям, были взяты пробы воды в 11 домохозяйствах. Получены следующие результаты (табл. 8.16). Проверка проводилась на 0,05-м уровне значимости. Поскольку представитель муниципалитета потребовал, чтобы содержание металлов было существенно меньше 40 частиц, примем эту величину в качестве генеральной медианы: $m = 40$; нулевая и альтернативная гипотезы будут записаны как

$$H_0: m \geqslant 40{,}0;$$
$$H_1: m \leqslant 40{,}0.$$

В табл. 8.16 приведены все необходимые данные. Для каждого домохозяйства определены $d_i = x_i - m_0$. Абсолютные значения этих разностей $|d_i|$ приведены в отдельной графе, затем ненулевые значения $|d_i|$ проранжированы. Наименьшее значение имеет 11 домохозяйство Л: $|d_{11}| = 0{,}9$, ему присваивается ранг 1, следующее значение: $|d_1| = 1{,}0$, это домохозяйство получает ранг 2. Домохозяйства Ж и И имеют одинаковые значения $|d_i|$, соответствующие рангам 3 и 4. Связанный ранг для них будет равен: $(3 + 4)/2 = 3{,}5$. Домохозяйство В имеет значение: $d_3 = 0$ и исключается из дальнейшего анализа. Для домохозяйств с $x_i > 40$ ранги вынесены в отдельную графу R^+. Сумма значений в этой графе дает статистику Вилкоксона: $W = 13$. Хотя значения графы R^- не используются, но эту графу полезно иметь, чтобы избежать ошибок.

Таблица 8.16

Расчет критерия Вилкоксона

Домохозяйство	Наблюдаемые значения, x_i	$d_i = x_i - m_0$	$\lvert d_i \rvert$	Ранг	R^+	R^-
А	39,0	−1,0	1	2		2
Б	20,2	−19,8	19,8	10		10
В	40,0	0,0	0,0	—		—
Г	32,2	−7,8	7,8	6		6
Д	30,5	−9,5	9,5	7		7
Е	26,5	−13,5	13,5	9		9
Ж	42,1	2,1	2,1	3,5	3,5	
З	45,6	5,6	5,6	5	5	
И	42,1	2,1	2,1	3,5	3,5	
К	29,9	−10,1	10,1	8		8
Л	40,9	0,9	0,9	1	1	
Итого	x	x	x	x	13	42

Для числа наблюдений с ненулевыми значениями $d_i n^*$ критерий знаков Вилкоксона будет находиться между $W = 0$ (если все d_i имеют отрицательные значения) и $W = \frac{n(n+1)}{2}$ (если все d_i имеют положительные значения). Наибольшее возможное значение $W = \frac{n(n+1)}{2}$ не что иное, как формула суммы последовательных чисел от 1 до n^2. Для 10 ненулевых значений W должно принимать значение в интервале [0, 55]. Заметим, что $\Sigma R^+ + \Sigma R^- = 55$.

Полученное значение $W = 13$ относительно низкое, принимая во внимание интервал возможных значений. Можно заключить, что питьевая вода скорее всего соответствует требованиям представителя муниципалитета.

Критическое значение критерия W может быть найдено из таблицы теста знаков Вилкоксона (табл. П.10 приложения). Приведены верхние и нижние значения для проверки левосторонней гипотезы. Для 10 ненулевых разностей $(x_i - m_0)$ и $\alpha = 0{,}05$ нижнее критическое значение W равно 11, верхнее — 44. Фактическое значение $W_{\text{факт}} = 13$ находится в интервале табличных значений критерия $W_{\text{табл}}$ [11÷44]. Следовательно, нулевая гипотеза не может быть отклонена на 5 %-ном уров-

не. Хотя выборочная проверка дает основания для беспокойства руководителям муниципалитета, все-таки для негативного вывода оснований недостаточно: на 5 %-ном уровне значимости можно заключить, что мнение о том, что питьевая вода не соответствует стандарту Комитета по здравоохранению, некорректно.

При большом числе наблюдений распределение W-статистики приближается к нормальному.

Формула, связывающая Z-статистику с критерием знаков Вилкоксона, следующая:

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}, \qquad (8.29)$$

где $W = \Sigma R^+$;

n — число наблюдений с $d_i \neq 0$.

Вычислим Z-статистику по данным нашего примера. Поскольку $n = 10$, то приближение W-статистики к Z будет более грубым, чем это было бы при $n \geqslant 20$. Подставляя $W = 13$, $n = 10$ в формулу (8.29), получаем $Z = -1{,}48$. Для 5 %-ного уровня критическое значение Z составляет $-1{,}645$ (по таблице нормального распределения). Поскольку фактическое значение находится справа от критической величины, нулевая гипотеза не может быть отклонена. Опасения руководителей муниципалитета о качестве воды необоснованны.

Критерий знаков Вилкоксона для сравнения двух выборок

Критерий знаков Вилкоксона может быть применен для сравнения двух выборок как непараметрический критерий решения той задачи, для которой ранее использовался параметрический t-критерий.

Обозначим характеристики одной совокупности x_i, а другой y_i. Совокупность величин $d_i = x_i - y_i$ предполагается симметричной или приблизительно симметричной, никаких предположений о характере распределения не делается. Поскольку методика применения критерия Вилкоксона для ре-

Таблица 8.17

Расчет критерия Вилкоксона при сравнении двух выборок

Номер задачи	Время, затраченное на решение задачи, с		$d_i = x_i - y_i$	$\lvert d_i \rvert$	Ранг	R^+	R^-
	x_i	y_i					
1	24,0	23,1	0,9	0,9	1	1	
2	16,7	20,4	−3,7	3,7	4		4
3	21,6	17,7	3,9	3,9	5	5	
4	23,7	20,7	3,0	3,0	2,5	2,5	
5	37,5	42,1	−4,6	4,6	6		6
6	31,4	36,1	−4,7	4,7	7		7
7	14,9	21,8	−6,9	6,9	10		10
8	37,3	40,3	−3,0	3,0	2,5		2,5
9	17,9	26,0	−8,1	8,1	11		11
10	15,5	15,5	0,0	0,0	—		—
11	29,0	35,4	−6,4	6,4	9		9
12	19,9	25,5	−5,6	5,6	8		8
Итого	x	x	x	x	x	8,5	55,5

шения этой задачи аналогична технике применения этого критерия к одной выборке, обратимся к примеру.

Пример. Фирма планирует приобрести пакет прикладных программ для решения задач финансового менеджмента. Из двух предложенных программных продуктов было решено выбрать тот, который обеспечивает большую скорость выполнения задач. Полученные результаты представлены в табл. 8.17.

Поскольку мы не знаем заранее, какой пакет обеспечивает большую скорость решения задач, нулевая и альтернативная гипотезы формулируются как H_0: $m_d = 0$, т. е. медиана величин $d_i = x_i - y_i$ в генеральной совокупности равна нулю; H_1: $m_d \neq 0$, т. е. медиана величин $d_i = x_i - y_i$ в генеральной совокупности не равна нулю.

В табл. 8.17 показаны разности $d_i = x_i - y_i$ для каждой пары наблюдений. Эти разности, взятые по абсолютной величине, проранжированы, причем ранг 1 присвоен минимальному значению $|d_i|$. Ранги положительных разностей показаны в графе R^-. Сумма рангов положительных разностей дает критерий Вилкоксона

$W = \Sigma R^+$.

Можно ли на 10%-ном уровне значимости сделать вывод о том, что медиана разности времени решения (m_d) для генеральной совокупности таких задач равна нулю?

По данным табл. 8.17 критерий Вилкоксона составил 8,5. Для 11 ненулевых значений разностей ($n = 11$) и $\alpha = 0{,}10$ критические значения критерия Вилкоксона — нижнее и верхнее равны 14 и 52 соответственно. Наблюдаемое значение $W = 8{,}5$ не попадает в эти пределы, следовательно, нулевая гипотеза отклоняется. Основываясь на имеющихся данных, можно сделать вывод, что сравниваемые программные продукты различаются по скорости выполнения задач. На 10 %-ном уровне значимости можно также заключить, что медиана $d_i = x_i - y_i$ меньше нуля, т. е. первый пакет прикладных программ обеспечивает более быстрое решение, чем второй пакет прикладных программ.

При $n \geqslant 20$ можно использовать Z-статистику. На основе $W = 8{,}5$ и $n = 11$ по формуле (8.29) вычисляем $Z = -2{,}18$. По таблице нормального распределения находим: $0{,}50000 - 0{,}4854 = 0{,}0146$ — область отклонения при левосторонней гипотезе. Ввиду того, что проверялась двусторонняя гипотеза, область отклонения испытуемой гипотезы такова: $0{,}0146 \cdot 2 = 0{,}0292$.

Критерий суммы рангов Вилкоксона для сравнения двух независимых выборок

Так же как и рассмотренный критерий знаков, критерий суммы рангов является непараметрическим, используемым для сравнения двух выборок, как и параметрический t-критерий. Этот тест предназначен для порядковых данных в предположении, что сравниваемые выборки — случайные и независимые. Никаких допущений о нормальности распределений и равенстве дисперсий не требуется.

Этот тест Вилкоксона эквивалентен тесту Манна—Уитни. Иногда его называют *двухвыборочным критерием Вилкоксона*. Предполагается, что *выборки различаются по объему*. Нулевая гипотеза может формулироваться как двусторонняя либо как односторонняя.

Таблица 8.18

Методика применения теста суммы рангов Вилкоксона

Гипотеза	Двусторонний тест	Левосторонний тест	Правосторонний тест
Нулевая, H_0	$m_1 = m_2$	$m_1 \geqslant m_2$	$m_1 \leqslant m_2$
Альтернативная, H_1	$m_1 \neq m_2$	$m_1 < m_2$	$m_1 > m_2$

Примечание. Здесь m_1 и m_2 — медианы генеральных совокупностей.

Меньшая выборка обычно обозначается номером 1. Если обе выборки равного размера, то номером 1 обозначается любая из них. Затем обе выборки объединяются и данные ранжируются в порядке возрастания, как если бы это была единая выборка. Наименьшее значение получает ранг 1, следующее — ранг 2 и т. д. Если значения совпадают, им присваиваются связанные ранги, которые определяются как средние из соответствующих порядковых номеров. Ранги для данных выборки 1 записываются в графу R_1, ранги для данных из выборки 2 записываются в графу R_2. Наблюдаемое (фактическое) значение критерия Вилкоксона рассчитывается по формуле

$$W = \Sigma R_1 \qquad (8.30)$$

Статистика U Манна—Уитни определяется как число пар (x_i, y_i), таких, что $x_i < y_i$ среди всех $n_1 n_2$ пар, в которых первый элемент — из первой выборки, второй — из второй. U и W связаны между собой следующей формулой:

$$U = \frac{n_1 n_2 + n_1(n_1 + 1)}{2} - W. \qquad (8.31)$$

Поскольку W и U линейно связаны, обычно говорят не о двух критериях, а об одном — двухвыборочном критерии Вилкоксона.

Критические значения теста W для сравнения двух выборок приведены в табл. П.11 приложения. Таблица критических значений содержит нижнее и верхнее критические значения критерия при соответствующих числах наблюдений в выборках n_1 и n_2. В случае больших объемов выборок может использоваться Z-статистика.

Таблица 8.19

Количество поворотов резиновых прокладок при испытании на эластичность

Номер партии	Резиновая прокладка									
	1	2	3	4	5	6	7	8	9	10
1	112	105	83	102	144	85	50			
2	91	183	141	219	105	138	146	86	134	106

Пример. Проверяя эластичность резиновых прокладок, контролер выбрал в случайном порядке семь изделий из одной партии и десять изделий из другой партии, которые были подвергнуты испытанию на эластичность. Фиксировалось количество поворотов каждого изделия до того, как резина оказывалась прорвана (табл. 8.19).

Можно ли считать на 5%-ном уровне значимости, что обе партии имеют одинаковую медиану количества оборотов изделий? Поскольку нет оснований полагать, что для изделий какой-то партии количество поворотов больше, чем для другой, нулевая и альтернативная гипотезы формулируются как двусторонние; H_0: $m_1 = m_2$ и H_1: $m_1 \neq m_2$, где m_1 и m_2 — медианы в первой и второй совокупностях.

Данные двух выборок объединяются и ранжируются. Наименьшему значению, 50 поворотов, присваивается ранг 1, следующему, 83 поворота, — ранг 2 и т. д. Наибольшему значению, 219 поворотов, присваивается ранг 17. Два одинаковых значения, 105, получили связанные ранги — 7,5 (табл. 8.20).

Итак, фактическое значение критерия W равно 44,5. Критическое значение критерия W находим по табл. П.11 приложения: для $n_1 = 7$ и $n_2 = 10$, $\alpha = 0{,}05$ $W_{\text{табл}} = (43;83)$, т. е. 43 — нижнее значение, 83 — верхнее значение. Поскольку фактическое значение $W = 44{,}5$ попадает в данный интервал, нулевая гипотеза не может быть отвергнута. Следовательно, с вероятностью 0,95 можно утверждать, что в обеих партиях медиана количества поворотов изделий при испытаниях на эластичность одинакова. Если $n_1 \geqslant 10$ и $n_2 \geqslant 10$, то может быть использовано нормальное распределение в качестве аппроксимации распределения критерия суммы рангов Вилкоксона. В этом случае рассчитывается Z-статистика по формуле

Таблица 8.20

Расчет двухвыборочного критерия Вилкоксона

Выборка 1	Ранг	Выборка 2	Ранг
112	10	91	5
105	7,5	183	16
83	2	141	13
102	6	219	17
144	14	105	7,5
85	4	138	12
50	1	146	15
$W = \Sigma R_1 = 44{,}5$		84	3
		134	11
		106	9
		$\Sigma R_2 = 108{,}5$	

$$Z = \frac{W - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}, \tag{8.32}$$

где $W = \Sigma R_1$; $n = n_1 + n_2$.

Полученное значение Z сравнивается с табличным значением z (по таблице нормального распределения при заданном α). Если $z_{\text{факт}} > z_{\text{табл}}$, H_0 отклоняется и наоборот.

Критерий Краскала—Уоллиса (H) для сравнения двух и более независимых выборок

Этот критерий используется для решения той же задачи, что и двухвыборочный критерий Вилкоксона, но при числе выборок больше двух. Нулевая гипотеза состоит в том, что медианы в генеральных совокупностях равны: H_0: $m_1 = m_2 = \ldots = m_k$. Альтернативная гипотеза H_1 заключается в том, что по крайней мере в одной из совокупностей медиана m_j принимает иное значение ($j = 1, 2, \ldots, k$).

Как и в случае применения критерия Вилкоксона, выборки объединяются, и данные ранжируются от минимального (ранг равен 1) до максимального (ранг равен: $n = n_1 + n_2 + \ldots + n_k$). При наличии связанных рангов они определяются как

средние значения из соответствующих рангов. По каждой выборке вычисляются квадраты суммы рангов: $(\Sigma R_1)^2$, ..., $(\Sigma R_k)^2$. Затем рассчитывается величина критерия H:

$$H = \frac{12}{n(n+1)}\left[\frac{(\Sigma R_1)^2}{n_1} + \frac{(\Sigma R_2)^2}{n_2} + \frac{(\Sigma R_3)^2}{n_3} + \ldots + \frac{(\Sigma R_k)^2}{n_k}\right] - 3(n+1). \quad (8.33)$$

Распределение статистики H аппроксимируется распределением хи-квадрата при условии, что каждая из выборок включает не менее пяти единиц; уровне значимости α и числе степеней свободы:

$$d.f. = k - 1,$$

где k — число сравниваемых совокупностей.

Если вычисленное значение статистики H превосходит $X^2_{\text{крит}}$, то нулевая гипотеза H_0 отклоняется.

Пример. Три малых предприятия с числом работников 6; 5 и 7 человек имеют одного и того же спонсора, который решил убедиться в том, что персонал во всех трех предприятиях имеет одну и ту же квалификацию. Было проведено тестирование работников, результаты которого представлены в табл. 8.21.

Рассчитаем тест Краскала—Уоллиса по данным табл. 8.21.

$$H = \frac{12}{18(18+1)}\left[\frac{32^2}{6} + \frac{45^2}{5} + \frac{94^2}{7}\right] - 3(18+1) = 7{,}512.$$

Таблица 8.21

Расчет критерия Краскала—Уоллиса

Оценка работника первого предприятия, баллов	Ранг	Оценка работника второго предприятия, баллов	Ранг	Оценка работника третьего предприятия, баллов	Ранг
67	7,5	64	5	75	15
57	1	73	13	61	3
62	4	72	12	76	16
59	2	68	9	71	11
70	10	65	6	78	17
67	7,5			74	14
				79	18
	$\Sigma R_1 = 32$		$\Sigma R_2 = 45$		$\Sigma R_3 = 94$

Критическое значение по таблице распределения хи-квадрат для $\alpha = 0{,}05$, $d.f. = 3 - 1 = 2$ равно 5,991 (табл. П.4 приложения). Поскольку 7,512 > 5,991, H_0 отклоняется, т. е. медианы квалификационных оценок работников различаются. Но если мы примем уровень значимости $\alpha = 0{,}01$, то этот вывод будет несостоятелен, поскольку $\chi^2 = 9{,}210$, $\alpha = 0{,}01$, $d.f. = 2$.

Основным непараметрическим критерием является критерий хи-квадрат. Важное значение имеет и непараметрический критерий Колмогорова—Смирнова. Непараметрические критерии занимают все более важное место в решениях задач статистического вывода, прежде всего с расширением анализа нечисловых данных (гл. 11).

РЕЗЮМЕ

Можно сделать статистический вывод — оценить свойства генеральной совокупности — с помощью испытания гипотез.

Процедура испытания всех гипотез одна и та же:

- определяем, что мы хотим узнать;
- формируем нулевую и альтернативную гипотезы;
- выбираем тестовую статистику (критерий);
- устанавливаем уровень значимости;
- вычисляем тестовую статистику (критерий) по данным выборки;
- находим критическое (табличное) значение критерия;
- сравниваем фактическое и критическое значения критерия и делаем вывод относительно нулевой гипотезы.

При испытании гипотезы о законе распределения используется непараметрический критерий: либо хи-квадрат Пирсона, либо критерий Колмогорова—Смирнова.

При испытании гипотезы о генеральной средней, если нам известна генеральная дисперсия σ^2, используется критерий $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.

Если мы не знаем σ^2, то оцениваем ее, используя выборочную дисперсию s^2, и проверочной статистикой является:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

с $(n-1)$ степенями свободы.

Для испытания одной выборочной доли p по сравнению с генеральной долей π, если объем выборки $n \leqslant 30$ и np или $n(1-p) \geqslant 5$, проверочная статистика рассчитывается следующим образом:

$$Z = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}}.$$

Ситуация усложняется, когда мы хотим испытать две независимые выборки.

Для испытания двух выборочных дисперсий s_1^2 и s_2^2 используем

$$F_{(n_1-1),(n_2-2)} = \frac{n_1 s_1^2}{(n_1-1)} \Big/ \frac{n_2 s_2^2}{(n_2-1)}$$

с большей дисперсией в числителе.

Для испытания выборочных средних $\bar{x}_1$ и $\bar{x}_2$, если мы знаем генеральные отклонения σ_1 и σ_2, статистическим критерием является:

$$t = \frac{|\bar{x}_1 - \bar{x}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Для того чтобы проверить две выборочные средние, если нам не известны генеральные дисперсии, мы должны сначала проверить гипотезу о равенстве генеральных дисперсий, используя F-критерий. Если F-критерий показывает, что мы можем предположить $\sigma_1^2 = \sigma_2^2$, то для проверки гипотезы о равенстве средних используется тестовая статистика:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 s_1^2 + n_2 s_2^2)}{(n_1 + n_2 - 2)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

с $(n_1 + n_2 - 2)$ степенями свободы.

Если F-критерий показывает, что мы должны принять $\sigma_1^2 \neq \sigma_2^2$, то при объемах выборок по крайней мере 30 наблюдений тестовой статистикой является:

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{(n_1-1)} + \frac{s_2^2}{(n_2-1)}}}.$$

Для испытания выборочных долей $\hat{p}_1$ и $\hat{p}_2$, если обе выборки не менее 30, тестовой статистикой является:

$$z = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

где $\bar{p}$ — доля из объединенных выборочных совокупностей.

Непараметрические критерии предпочтительны, поскольку не требуют предположений о характере распределения генеральной совокупности. Все чаще используется критерий знаков Вилкоксона, который применяется как к данным одной выборки, так и к данным двух сравнимых выборок. Для сравнения двух неравных выборок в случае порядковых данных может использоваться критерий суммы рангов Вилкоксона; для сравнения более двух выборок используется непараметрический критерий Краскала—Уоллиса.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Айвазян С. А., Мхитарян В. С.* Теория вероятностей и прикладная статистика. Т. 1: Учебник для вузов. — М.: ЮНИТИ, 2001.
2. *Закс Л.* Статистическое оценивание: Пер. с нем. / Под ред. и с предисл. Ю. П. Адлера и В. Г. Горского. — М.: Статистика, 1976.
3. *Козлов А. Ю., Шишлов В. Ф.* Пакет анализа MS Excel в экономико-статистических расчетах / Под ред. В. С. Мхитаряна. — М.: ЮНИТИ — ДАНА, 2003.
4. *Ниворожкина Л. И., Морозова З. А.* Сборник задач по математической статистике с элементами теории вероятностей РИНХ. — Ростов-на-Дону, 2002.
5. *Эддоус М., Стэнсфилд Р.* Методы принятия решений: Пер. с англ. / Под ред. И. И. Елисеевой. — М.: ЮНИТИ, 1997.

Глава
9

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ И МОДЕЛИРОВАНИЕ СТАТИСТИЧЕСКИХ СВЯЗЕЙ

9.1. Понятие о статистической и корреляционной связи

Современная наука исходит из взаимосвязей всех явлений природы и общества. Объем продукции предприятия связан с численностью работников, мощностью двигателей, стоимостью производственных фондов и еще многими признаками.

Невозможно управлять явлениями, предсказывать их развитие без изучения характера, силы и других особенностей связей. Поэтому методы исследования, измерения связей составляют чрезвычайно важную часть методологии научного исследования, в том числе и статистического.

Различают два типа связей между различными явлениями и их признаками: *функциональную,* или *жестко детерминированную,* с одной стороны, и *статистическую,* или *стохастически детерминированную,* — с другой. Строго определить различие этих типов связи можно тогда, когда они получают математическую формулировку. Для простоты будем говорить о связи двух явлений или двух признаков, математически отображаемой в форме уравнения связи двух переменных.

Если с изменением значения одной из переменных вторая изменяется строго определенным образом, т.е. значению одной переменной обязательно соответствует одно или несколько точно заданных значений другой переменной, связь между ними является *функциональной.*

Нередко говорят о строгом соответствии лишь одного значения второй из переменных каждому значению первой из них, но это неверно. Например, связь между y и x является строго функциональной, если $y = \sqrt{x}$; но значению $x = 4$ соответствует не одно, а два значения: $y_1 = +2$, $y_2 = -2$. Уравнения более высоких степеней могут иметь несколько корней, связь, разумеется остается функциональной.

Функциональная связь двух величин возможна лишь при условии, что вторая из них зависит только от первой. В реальной природе (и тем более в обществе) таких связей нет; они являются лишь абстракциями, полезными и необходимыми при анализе явлений, но упрощающими реальность. Функциональная зависимость данной величины y от многих факторов $x_1, x_2, \ldots, x_k$ возможна только в том случае, если величина y всегда зависит только от перечисленного набора факторов $x_1, x_2, \ldots, x_k$ и ни от чего более. Все явления и процессы реального мира связаны между собой, и нет такого конечного числа переменных k, которое абсолютно полно определяло бы собой зависимую величину y. Следовательно, множественная функциональная зависимость переменных есть тоже абстракция, упрощающая реальность.

Однако механика, электротехника, акустика, политическая экономия и другие науки успешно используют представление связей как функциональных не только в аналитических целях, но нередко и в целях прогнозирования. Это возможно потому, что в простых системах интересующая нас переменная величина зависит в основном (скажем, на 99% или даже на 99,99%) от немногих других переменных или только от одной переменной, т.е. связь является хотя и не абсолютно функциональной, но практически очень близкой к таковой. Например, длина года (период обращения Земли вокруг Солнца) почти функционально зависит только от массы Солнца и расстояния Земли от него. На самом деле она зависит в очень слабой степени и от масс, и расстояния других планет от Земли, но вносимые ими (и тем более далекими звездами) искажения функциональной связи для всех практических целей, кроме космонавтики, пренебрежимо малы.

Стохастически детерминированная связь не имеет ограничений и условий, присущих функциональной связи. Если с изменением значения одной из переменных вторая может в определенных пределах принимать любые значения с некоторыми вероятностями, но ее среднее значение или иные статистические (массовые) характеристики изменяются по определенному закону, связь является статистической. Иными словами, при статистической связи разным значениям одной переменной соответствуют разные распределения значений другой переменной.

В настоящее время наука не знает более широкого определения связи. Все связи, которые могут быть измерены и выражены численно, подходят под определение «статистические связи», в том числе и функциональные. Последние представляют собой частный случай статистических связей, когда значениям одной переменной соответствуют «распределения» значений второй, состоящие из одного или нескольких значений и имеющие вероятность, равную единице. Конечно, качественное различие действительно вероятностных распределений и отдельных значений, имеющих вероятность единицы (достоверных), настолько велико, что хотя функциональные связи и могут рассматриваться как предельный случай статистической связи, все же с полным основанием можно говорить о двух типах связей.

Корреляционной связью называют важнейший частный случай статистической связи, состоящий в том, что разным значениям одной переменной соответствуют различные средние значения другой. С изменением значения признака x закономерным образом изменяется среднее значение признака y, в то время как в каждом отдельном случае значение признака y (с различными вероятностями) может принимать множество различных значений.

Если же с изменением значения признака x среднее значение признака y не изменяется закономерным образом, но закономерно изменяется другая статистическая характеристика (показатели вариации, асимметрии, эксцесса и т.п.), то связь не является корреляционной, но статистической.

Статистическая связь между двумя признаками (переменными величинами) предполагает, что каждый из них имеет случайную вариацию индивидуальных значений относительно средней величины. Если же такую вариацию имеет только один из признаков, а значения другого являются жестко детерминированными, то говорят лишь о регрессии. Например, при анализе динамических рядов можно измерять регрессию уровней ряда урожайности (имеющих случайную колеблемость) на номера лет. Но нельзя говорить о корреляции между ними и применять показатели корреляции с соответствующей интерпретацией (гл. 10).

Само слово *корреляция* ввел в статистику английский биолог и статистик Френсис Гальтон в конце XIX в. Тогда оно писалось как «*correlation*» (соответствие), но не просто «связь» (*relation*), а «как бы связь», т.е. связь, но не в привычной в то время функциональной форме. В науке вообще, а именно в палеонтологии, термин «корреляция» применил еще раньше, в конце XVIII в., знаменитый французский палеонтолог (специалист по ископаемым останкам животных и растений прошлых эпох) Жорж Кювье. Он ввел даже «закон корреляции» частей и органов животных. «Закон корреляции» помогает восстановить по найденным в раскопках черепу, костям и т.д. облик всего животного и его место в системе: если череп с рогами, то это было травоядное животное, а его конечности имели копыта; если же лапы с когтями — то хищное животное без рогов, но с крупными клыками.

Известен следующий рассказ о Кювье и «законе корреляции». В дни университетского праздника студенты решили подшутить над профессором Кювье. Они вырядили одного из студентов в козлиную шкуру с рогами и копытами и подсадили его в окно спальни Кювье. Ряженый затопал копытами и завопил: «Я тебя съем!» Кювье проснулся, увидел силуэт с рогами и спокойно отвечал: «Если у тебя рога и копыта, то по закону корреляции ты травоядное, и съесть меня не можешь. А за то, что не знаешь закона корреляции, получишь двойку!»

Корреляционная связь между признаками может возникнуть разными путями. Первый (важнейший) путь — *причинная зависимость* результативного признака (его вариации) от вариации факторного признака. Например, признак x — балл оценки плодородия почв, признак y — урожайность сельскохозяйственной культуры. Здесь совершенно ясно логически, какой признак выступает как независимая переменная (фактор) x, какой — как зависимая переменная (результат) y.

Второй путь — *сопряженность*, возникающая при наличии общей причины. Известен классический пример, приведенный крупнейшим статистиком России начала XX в. А. А. Чупровым: если в качестве признака x взять число пожарных команд в городе, а за признак y — сумму убытков за год в городе от пожаров, то между признаками x и y в совокупности городов России существовала прямая корреляция; в среднем чем больше пожарников в городе, тем больше и убытков от пожаров! Уж не занимались ли пожарники поджигательством из боязни потерять работу? Но дело в другом. Данную корреляцию нельзя интерпретировать как связь причины и следствия; оба признака-следствия общей причины — размера города. Вполне логично, что в крупных городах больше пожарных частей, но больше и пожаров, и убытков от них за год, чем в малых городах.

Третий путь возникновения корреляции — *взаимосвязь признаков*, каждый из которых и причина, и следствие. Такова, например, корреляция между уровнями производительности труда рабочих и уровнем оплаты 1 ч труда (тарифной ставкой). С одной стороны, уровень зарплаты — следствие производительности труда: чем она выше, тем выше и оплата. Но, с другой стороны, установленные тарифные ставки и расценки играют стимулирующую роль: при правильной системе оплаты они выступают в качестве фактора, от которого зависит производительность труда. В такой системе признаков допустимы обе постановки задачи; каждый признак может выступать в роли независимой переменной x и в качестве зависимой переменной y.

9.2. Условия применения и ограничения корреляционно-регрессионного метода

Поскольку корреляционная связь является статистической, первым условием возможности ее изучения является наличие данных по достаточно большой совокупности. По отдельным явлениям можно получить совершенно превратное представление о связи признаков, ибо в каждом отдельном явлении значения признаков, кроме закономерной составляющей, имеют случайное отклонение (вариацию). Например, сравнивая два хозяйства, одно из которых

имеет лучшее качество почв, по уровню урожайности, можно обнаружить, что урожайность выше в хозяйстве с худшими почвами. Ведь урожайность зависит от сотен факторов и при том же самом качестве почв может быть и выше, и ниже. Но если сравнивать большое число хозяйств с лучшими почвами и большое число — с худшими, то средняя урожайность в первой группе окажется выше и станет возможным измерить достаточно точно параметры корреляционной связи.

Какое именно число явлений достаточно для анализа корреляционной и вообще статистической связи, зависит от цели анализа, требуемой точности и надежности параметров связи, от числа факторов, корреляция с которыми изучается. Обычно считают, что число наблюдений должно быть не менее чем в 5—6, а лучше — в 10 раз больше числа факторов. Еще лучше, если число наблюдений в несколько десятков или в сотни раз больше числа факторов, тогда закон больших чисел обеспечивает эффективное взаимопогашение случайных отклонений от закономерного характера связи признаков.

Вторым условием закономерного проявления корреляционной связи служит условие, обеспечивающее надежное выражение закономерности в средней величине. Кроме уже указанного большого числа единиц совокупности для этого необходима достаточная однородность совокупности. Нарушение этого условия может извратить параметры корреляции. Например, в массе зерновых хозяйств уровень продукции с 1 га растет по мере концентрации площадей, т.е. он выше в крупных хозяйствах. В массе овощных и овоще-молочных хозяйств (пригородный тип) наблюдается та же прямая связь уровня продукции с размером хозяйства. Но если соединить в общую неоднородную совокупность те и другие хозяйства, то связь уровня продукции с размером площади пашни (или посевной площади) получится обратной. Причина в том, что овощные и овоще-молочные хозяйства, имея меньшую площадь, чем зерновые, производят больше продукции с 1 га ввиду большей интенсивности производства в данных отраслях.

В качестве третьего условия корреляционного анализа выдвигается необходимость подчинения распределения совокупности по результативному и факторным признакам нормальному закону распределения вероятностей. Это усло-

вие связано с применением метода наименьших квадратов при расчете параметров корреляции: только при нормальном распределении метод наименьших квадратов дает оценки параметров, отвечающих принципам максимального правдоподобия. На практике эта предпосылка чаще всего выполняется приближенно, но и тогда метод наименьших квадратов дает неплохие результаты[1].

Однако при значительном отклонении распределений признаков от нормального закона нельзя оценивать надежность выборочного коэффициента корреляции, используя параметры нормального распределения вероятностей или распределения Стьюдента.

Еще одним спорным вопросом является допустимость применения корреляционного анализа к функционально связанным признакам. Можно ли, например, построить уравнение корреляционной зависимости размеров выручки от продажи картофеля, от объема продажи и цены? Ведь произведение объема продажи и цены равно выручке в каждом отдельном случае. Как правило, к таким жестко детерминированным связям применяют только индексный метод анализа. Однако на этот вопрос можно взглянуть и с другой точки зрения. При индексном анализе выручки предполагается, что количество проданного картофеля и его цена независимы друг от друга, потому-то и допустима абстракция от изменения одного фактора при изменении влияния другого, как это принято в индексном методе (гл. 13). В реальности количество и цена не являются вполне независимыми друг от друга.

Возможные связи в системе трех переменных представлены на рис. 9.1.

Корреляционно-регрессионный анализ учитывает межфакторные связи, следовательно, дает более полное измерение роли каждого фактора: прямое, непосредственное его влияние на результативный признак; косвенное влияние фактора через его влияние на другие факторы; влияние всех факторов на результативный признак. Если связь между факторами несущественна, можно ограничиться индексным анали-

[1] *Крастинь О. П.* Разработка и интерпретация моделей корреляционных связей в экономике. — Рига: Зинатне, 1983. — С. 14.

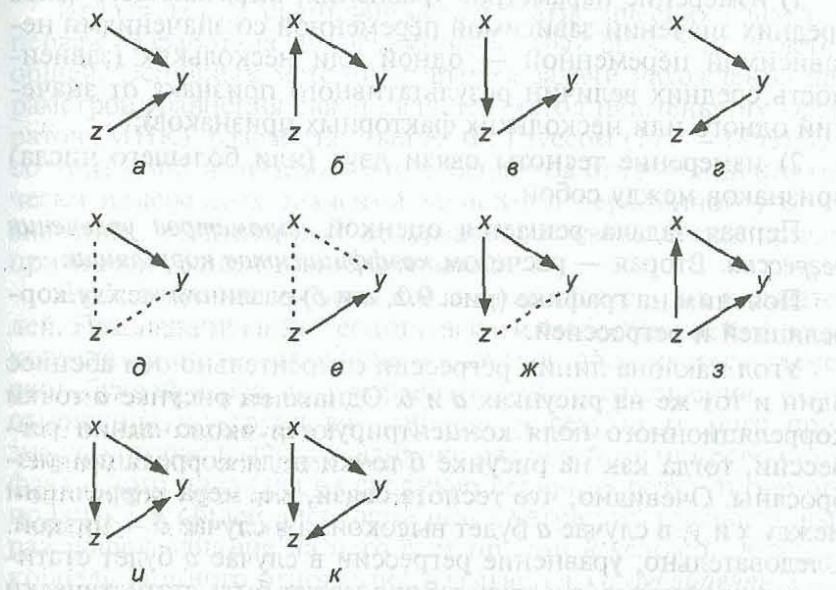


Рис. 9.1. Связи в системе трех переменных:
а — обе переменные *x* и *z* влияют на *y*; *б* — переменная *z* не влияет на *y*; ее влияние полностью входит в *x*; *в* — переменная *z* поглощает влияние *x* и передает его, влияя на *y*; *г* — переменная *z* — субследствие из *y*; *д* — переменная *z* не влияет на *y*; *е* — переменная *x* не влияет на *y*; *ж* — переменные *z* и *y* не связаны между собой, но имеют общую причину — *x* (классический случай «ложной корреляции»); *з* — переменная *z* передает свое влияние на *y* как непосредственно, так и через *x*; *и* — переменная *x* влияет на *y* как непосредственно, так и через *z*; *к* — переменная *x* влияет как на *z*, так и на *y* и конкурирует с *y* во влиянии на *z*

зом. В противном случае его полезно дополнить корреляционно-регрессионным измерением влияния факторов, даже если они функционально связаны с результативным признаком.

9.3. Задачи корреляционно-регрессионного анализа и моделирования

В соответствии с сущностью корреляционной связи ее изучение имеет две задачи:

1) измерение параметров уравнения, выражающего связь средних значений зависимой переменной со значениями независимой переменной — одной или нескольких (зависимость средних величин результативного признака от значений одного или нескольких факторных признаков);

2) измерение тесноты связи двух (или большего числа) признаков между собой.

Первая задача решается оценкой *параметров уравнения регрессии*. Вторая — расчетом *коэффициентов корреляции*.

Поясним на графике (рис. 9.2, *а* и *б*) различия между корреляцией и регрессией.

Угол наклона линии регрессии относительно оси абсцисс один и тот же на рисунках *а* и *б*. Однако на рисунке *а* точки корреляционного поля концентрируются около линии регрессии, тогда как на рисунке *б* точки поля корреляции разбросаны. Очевидно, что теснота связи, т.е. мера корреляции между x и y, в случае *а* будет высокой, а в случае *б* — низкой. Следовательно, уравнение регрессии в случае *а* будет статистически значимо, а в случае *б* оно может быть статистически незначимо. Таким образом, случаи *а* и *б* различаются величиной коэффициентов корреляции, но в то же время будут иметь одинаковые коэффициенты регрессии:

$$(а)\ r_{yx} \neq (б)\ r_{yx};$$
$$(а)\ b_{yx} = (б)\ b_{yx}.$$

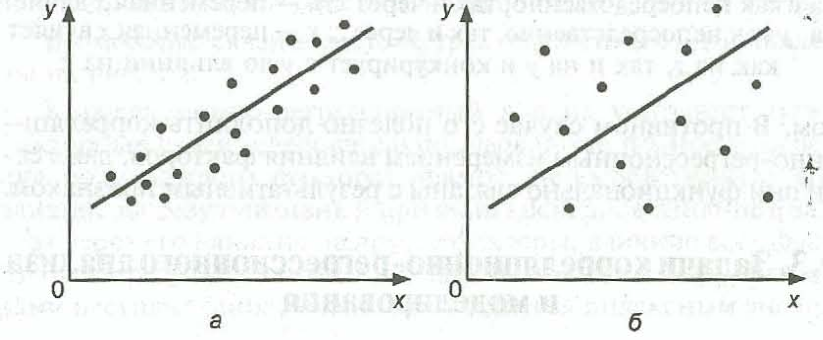


Рис. 9.2. Регрессия при разной интенсивности корреляции: *а* — тесная корреляция; *б* — слабая корреляция

Вторая задача специфична для статистических связей, а первая разработана для функциональных связей и является общей. Основным методом решения задачи нахождения параметров уравнения связи является метод наименьших квадратов (МНК), разработанный К. Ф. Гауссом (1777—1855). Он состоит в минимизации суммы квадратов отклонений фактически измеренных значений зависимой переменной y от ее значений, вычисленных по уравнению связи с факторным признаком, одним или несколькими, x.

Для измерения тесноты связи применяется ряд показателей. При парной связи теснота связи измеряется прежде всего корреляционным отношением, которое обозначается греческой буквой η. Квадрат корреляционного отношения — это отношение межгрупповой дисперсии результативного признака, которая выражает влияние различий группировочного факторного признака на среднюю величину результативного признака, к общей дисперсии результативного признака, выражающей влияние на него всех причин и условий. Квадрат корреляционного отношения называется *коэффициентом детерминации*:

$$\eta^2 = \frac{\sum_{j=1}^{k} (\bar{y}_j - \bar{y})^2 \cdot f_j}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \qquad (9.1)$$

где k — число групп по факторному признаку;
$\bar{y}$ — общее среднее значение;
f_j — частота в j-й группе;
n — число единиц в совокупности;
y_i — значение результативного признака для i-й единицы;
$\bar{y}_j$ — среднее значение y в j-й группе.

Формула (9.1) используется при расчете показателя тесноты связи по аналитической группировке (см. гл. 6). При вычислении корреляционного отношения по уравнению регрессии, парному или множественному, применяется формула (9.2):

$$\eta^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \qquad (9.2)$$

где $\hat{y}_i$ — значение y для i-й единицы, рассчитанное по уравнению регрессии.

Сумма квадратов в числителе — это дисперсия результативного признака y, объясненная связью с фактором x (факторами). Она вычисляется по индивидуальным данным, полученным для каждой единицы совокупности на основе уравнения регрессии, и называется дисперсией, *объясненной уравнением регрессии*.

Если уравнение выбрано неверно или сделана ошибка при расчете его параметров, то сумма квадратов в числителе может оказаться большей, чем в знаменателе, и отношение утратит тот смысл, который оно должно иметь, а именно: какова доля общей вариации результативного признака, объясняемая на основе выбранного уравнения связи его с факторным признаком (признаками). Чтобы избежать ошибочного результата, лучше вычислять корреляционное отношение по другой формуле (9.3), не столь наглядно выявляющей сущность показателя, но зато полностью гарантирующей от возможного искажения:

$$\eta^2 = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}. \qquad (9.3)$$

В числителе формулы (9.3) стоит сумма квадратов отклонений фактических значений признака y от его индивидуальных расчетных значений $\hat{y}_i$, т.е. вся дробь — это доля вариации результативного признака, не объясненная входящими в уравнение связи признаками-факторами. Эта сумма не может быть равной нулю, если связь не является функциональной. При неверной формуле уравнения связи или ошибке в расчетах возрастает расхождение фактических и расчетных значений, и корреляционное отношение снижается, как логически и должно быть.

В основе перехода от формулы (9.2) к формуле (9.3) лежит известное правило разложения сумм квадратов отклонений:

$$D_{\text{общ}} = D_{\text{межгр}} + D_{\text{внутригр}}.$$

Согласно этому правилу можно вместо межгрупповой (факторной) дисперсии использовать разность

$$D_{\text{межгр}} = D_{\text{общ}} - D_{\text{внутригр}},$$

что дает:

$$\eta^2 = \frac{D_{\text{общ}} - D_{\text{внутригр}}}{D_{\text{общ}}} = 1 - \frac{D_{\text{внутригр}}}{D_{\text{общ}}}.$$

При расчете η не по группировке, а по уравнению корреляционной связи (уравнению регрессии) мы используем формулу (9.3). В этом случае правило разложения суммы квадратов отклонений результативного признака записывается как

$$D_{\text{общ}} = D_{\text{объясн}} + D_{\text{ост}}.$$

Важнейшее положение, которое следует теперь усвоить любому желающему правильно применять методы корреляционно-регрессионного анализа, состоит в интерпретации формул (9.2) и (9.3) и гласит.

Уравнение корреляционной связи измеряет зависимость между вариаций результативного признака и вариацией факторного признака (признаков). Меры тесноты связи измеряют долю вариации результативного признака, которая связана корреляционно с вариацией факторного признака (признаков).

Интерпретировать корреляционные показатели следует строго в терминах вариации (различий в пространстве) отклонений от средней величины. Если же задача исследования состоит в измерении связи не между вариацией двух признаков в совокупности, а между изменениями признаков объекта во времени, то метод корреляционно-регрессионного анализа требует значительного изменения (гл. 12).

Из вышеприведенного положения об интерпретации показателей корреляции следует, что нельзя трактовать корреляцию признаков как причинную связь их уровней.

Пример. Если бы все крестьяне области внесли под картофель одинаковую дозу удобрений, то вариация этой дозы была бы равна нулю, а следовательно, она абсолютно не могла бы влиять на вариацию урожайности картофеля. Параметры корреляции дозы удобрений с урожайностью будут тогда строго равны нулю. Но ведь и в этом случае уровень урожайности зависел бы от дозы удобрений — он был бы выше, чем без удобрений.

Итак, строго говоря, метод корреляционно-регрессионного анализа не может объяснить роли факторных признаков в создании результативного признака. Это очень серьезное ограничение метода, о котором не следует забывать.

Следующий общий вопрос — это уже рассмотренный в разделе о группировке вопрос о «чистоте» измерения влияния каждого отдельного факторного признака. Как отмечалось в главе 6, группировка совокупности по одному факторному признаку может отразить влияние именно данного фактора на результативный признак при условии, что все другие факторы не связаны с изучаемым, а случайные отклонения и ошибки взаимопогасились в большой совокупности. Если же изучаемый фактор связан с другими факторами, влияющими на результативный признак, будет получена не «чистая» характеристика влияния только одного фактора, а сложный комплекс, состоящий как из непосредственного влияния фактора, так и из его косвенных влияний, через его связь с другими факторами и их влияние на результативный признак. Данное положение полностью относится и к парной корреляционной связи.

Однако коренное отличие метода корреляционно-регрессионного анализа от аналитической группировки состоит в том, что корреляционно-регрессионный анализ позволяет разделить влияние комплекса факторных признаков, анализировать различные стороны сложной системы взаимосвязей. Если метод комбинированной аналитической группировки, как правило, не дает возможность анализировать более трех факторов, то корреляционный метод при объеме совокупности около 100 единиц позволяет вести анализ системы с 8—10 факторами и разделить их влияние.

Наконец, развивающиеся на базе корреляционно-регрессионного анализа многомерные методы (метод главных компонент, факторный анализ) позволяют синтезировать влияние признаков (наблюдаемых факторов), выделяя из них непосредственно неучитываемые глубинные факторы (компоненты). Например, изучая корреляцию ряда признаков интенсификации сельскохозяйственного производства, таких, как фондообеспеченность, затраты труда на единицу площади, энергообеспеченность, внесение удобрений на единицу площади, плотность поголовья скота, можно синтезировать

их влияние на уровень продукции с единицы площади, или на производительность труда, получив обобщенный фактор «интенсификация производства», непосредственно неизмеримый.

Правильное применение и интерпретация результатов корреляционно-регрессионного анализа возможны лишь при понимании всех специфических черт, достоинств и ограничений метода. Поэтому рекомендуем вернуться к данному подразделу заново после изучения остальных разделов этой главы и после приобретения некоторой практики применения метода к решению различных задач.

Необходимо сказать и о других задачах, решаемых с помощью корреляционно-регрессионного метода, имеющих не формально математический, а содержательный характер.

1. *Задача выделения важнейших факторов*, влияющих на результативный признак (т.е. на вариацию его значений в совокупности). Эта задача решается в основном на базе мер тесноты связи признаков-факторов с результативным признаком.

2. *Задача оценки хозяйственной деятельности по эффективности использования имеющихся факторов производства.* Эта задача решается путем расчета для каждой единицы совокупности тех величин результативного признака, которые были бы получены при средней по совокупности эффективности использования факторов в сравнении их с фактическими результатами производства.

3. *Задача прогнозирования возможных значений результативного признака при задаваемых значениях факторных признаков.*

Такая задача решается путем подстановки ожидаемых, или планируемых, или возможных значений факторных признаков в уравнение связи и вычисления ожидаемых значений результативного признака.

Приходится решать и обратную задачу: вычисление необходимых значений факторных признаков для обеспечения планового, или желаемого, значения результативного признака в среднем по совокупности. Эта задача обычно не имеет единственного решения в рамках данного метода и должна дополняться постановкой и решением оптимизационной задачи на нахождение наилучшего из возможных вариантов ее решения (например, варианта, позволяющего достичь требуемого результата с минимальными затратами).

4. *Задача подготовки данных, необходимых в качестве исходных для решения оптимизационных задач.* Например, для нахождения оптимальной структуры производства в районе на перспективу исходная информация должна включать показатели производительности на предприятиях разных отраслей и форм собственности. В свою очередь, эти показатели могут быть получены на основе корреляционно-регрессионной модели либо на основе тренда динамического ряда (а тренд — это тоже уравнение регрессии).

При решении каждой из названных задач нужно учитывать особенности и ограничения корреляционно-регрессионного метода. Всякий раз необходимо специально обосновать возможность причинной интерпретации уравнения как *объясняющего* связь между вариацией фактора и результата. Трудно обеспечить раздельную оценку влияния каждого из факторов. В этом отношении корреляционные методы глубоко противоречивы. С одной стороны, их идеал — измерение чистого влияния каждого фактора. С другой стороны, такое измерение возможно при отсутствии связи между факторами и случайной вариацией признаков. А тогда связь является функциональной, и корреляционные методы анализа излишни. В реальных системах связь всегда имеет статистический характер, и тогда идеал методов корреляции становится недостижимым. Но это не значит, что данные методы не нужны.

Указанное противоречие означает попросту недостижимость абсолютной истины в познании реальных связей. Приближенный характер любых результатов корреляционно-регрессионного анализа не является поводом для отрицания их полезности. Любая научная истина — относительна. Забыть об этом и абсолютизировать параметры регрессионных уравнений, меры корреляции было бы ошибкой, так же как и отказаться от использования этих мер.

9.4. Вычисление и интерпретация параметров парной линейной регрессии

Простейшей системой корреляционной связи является линейная связь между двумя признаками — *парная линейная корреляция.*

Практическое ее значение в том, что есть системы, в которых среди всех факторов, влияющих на результативный признак, выделяется один важнейший фактор, который в основном определяет вариацию результативного признака. Измерение парных корреляций составляет необходимый этап в изучении сложных, многофакторных связей. Есть и такие системы связей, при изучении которых следует предпочесть парную корреляцию. Внимание к линейным связям объясняется ограниченной вариацией переменных и тем, что в большинстве случаев нелинейные формы связей для выполнения расчетов преобразуются в линейную форму (линеаризуются).

Уравнение парной линейной корреляционной связи называется *уравнением парной регрессии* и имеет вид:

$$\tilde{y} = a + bx, \qquad (9.4)$$

где $\tilde{y}$ — среднее значение результативного признака y при определенном значении факторного признака x;

a — свободный член уравнения;

b — коэффициент регрессии, измеряющий среднее отношение отклонения результативного признака от его средней величины к отклонению факторного признака от его средней величины на одну единицу его измерения, — вариация y, приходящаяся на единицу вариации x.

Что касается термина «регрессия», его происхождение таково: создатели корреляционного анализа Ф. Гальтон (1822—1911) и К. Пирсон (1857—1936) интересовались связью между ростом отцов и их сыновей. Ф. Гальтон изучил более 200 семей и обнаружил, что в группе семей с высокорослыми отцами сыновья в среднем ниже ростом, чем их отцы, а в группе семей с низкорослыми отцами сыновья в среднем выше отцов. Таким образом, отклонение роста от средней в следующем поколении уменьшается — *регрессирует.* Причина в том, что на рост сыновей влияет не только рост отцов, но и рост матерей и много других факторов развития ребенка, и эти факторы, случайно направленные как в сторону увеличения, так и снижения роста, конечно, приближают рост сыновей к среднему росту. В целом же вариация роста, конечно, не уменьшается, а в наше время «акселерации» сам средний рост увеличивается из поколения в поколение (до известного предела).

Параметры уравнения (9.4) рассчитываются методом наименьших квадратов (МНК) по данным о значениях признаков x и y в изучаемой совокупности, состоящей из n единиц.

Исходное условие МНК для прямой линии имеет вид:

$$f(a, b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \to \min. \qquad (9.5)$$

Для отыскания значений параметров a и b, при которых $f(a, b)$ принимает минимальное значение, частные производные функции приравниваем нулю и преобразуем полученные уравнения, которые называются *нормальными уравнениями МНК для прямой:*

$$\frac{\partial f}{\partial a} = 2\sum_{i=1}^{n}(y_i - a - bx_i)(-1) = 0;$$

$$\frac{\partial f}{\partial b} = 2\sum_{i=1}^{n}(y_i - a - bx_i)(-x_i) = 0.$$

Отсюда система нормальных уравнений имеет вид:

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i;$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Нормальные уравнения (МНК) для прямой линии регрессии являются системой двух уравнений с двумя неизвестными a и b. Все остальные величины, входящие в систему, определяются по исходной информации. Таким образом, оба параметра уравнения линейной регрессии однозначно вычисляются при решении этой системы уравнений.

Если первое нормальное уравнение разделить на n, получим:

$$a + b\bar{x} = \bar{y}, \text{ откуда } a = \bar{y} - b\bar{x}. \qquad (9.6)$$

По уравнению (9.6) обычно на практике вычисляется свободный член уравнения регрессии a. Параметр b — *коэффициента регрессии* вычисляется по преобразованной формуле, которую можно вывести, решая систему нормальных уравнений относительно b:

$$b = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}. \qquad (9.7)$$

Так как знаменатель этого выражения есть не что иное, как дисперсия признака, т.е. σ_x^2, то можно записать формулу коэффициента регрессии в виде

$$b = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x^2}. \qquad (9.8)$$

Подставив в (9.8) выражение σ_x^2, получим:

$$b = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]:n} = \frac{n\overline{xy} - n\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (9.9)$$

В числителе (9.9) ковариация переменных x и y.

Параметры уравнения регрессии можно вычислить через определители:

$$a = \frac{\Delta_a}{\Delta}, \qquad b = \frac{\Delta_b}{\Delta}, \qquad (9.10)$$

где Δ — определитель системы;

Δ_a — частный определитель, получаемый в результате замены коэффициентов при a свободными членами из правой части системы уравнений;

Δ_b — частный определитель, получаемый в результате замены коэффициентов при b свободными членами из правой части системы уравнений.

Формулы (9.10) соответствуют самому общему подходу к определению параметров уравнения регрессии и могут применяться в случае как парной, так и множественной регрессии.

Применение одной из формул (9.7), (9.8) или (9.9) зависит от характера данных и наличия уже вычисленных на предыдущих этапах анализа показателей. Если были вычислены x, y,

σ_x, σ_y, то проще применить формулу (9.7) или (9.8). Если расчет параметров управления ведется по первичным данным x_i, y_i, то удобнее формула (9.9). Ее использование существенно сокращает объем вычислений при слабой вариации признаков, ибо тогда отклонения их индивидуальных значений от средних величин на порядок или два меньше самих индивидуальных и средних величин. Помимо того формула (9.9) явно выражает указанную в подразд. 9.1 особенность корреляционного анализа связей: параметры корреляции зависят не от уровней признаков, а только от их отклонений от средних значений.

Если значения признака увеличить в 10 раз, корреляция не изменится; также не изменятся параметры уравнения, кроме свободного члена, если ко всем значениям каждого признака прибавить постоянное число.

Коэффициент парной линейной регрессии, обозначенный b, имеет смысл показателя *силы связи* между вариацией факторного признака x и вариацией результативного признака y. Он измеряет среднее по совокупности отклонение y от его средней величины при отклонении признака x от своей средней величины на принятую единицу измерения.

Например, по данным табл. 9.1, при отклонении затрат на одну корову от средней величины на 1 руб. надой молока на корову отклоняется от своего среднего значения на 3,47 кг в среднем по совокупности. При отклонении фактора на $x_i - \bar{x}$ результативный признак отклоняется в среднем на $\tilde{y}_i - \bar{y}$.

Теснота парной линейной корреляционной связи, как и любой другой формы связи, может быть измерена корреляционным отношением η. Кроме того, при линейной форме уравнения применяется другой показатель тесноты связи — коэффициент корреляции r_{yx}. Этот показатель представляет собой *стандартизованный коэффициент регрессии*, т.е. коэффициент, выраженный не в абсолютных единицах измерения признаков, а в долях среднего квадратического отклонения результативного признака:

$$r_{yx} = b\frac{\sigma_x}{\sigma_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 : n}}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2 : n}} =$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (9.11)$$

или

$$r_{yx} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}. \qquad (9.12)$$

Коэффициент корреляции может принимать значения $-1 \leqslant r \leqslant 1$; по абсолютной величине $0 \leqslant |r| \leqslant 1$. Отрицательные значения r_{yx} свидетельствуют об обратной связи признаков y и x, положительные — о прямой связи. Коэффициент корреляции симметричная мера: $r_{yx} = r_{xy}$.

Коэффициент корреляции был предложен английским статистиком и философом Карлом Пирсоном. В соответствии с формулой (9.12) интерпретация r_{yx} такова: отклонение признака-фактора от его среднего значения на величину своего среднего квадратического отклонения в среднем по совокупности приводит к отклонению признака-результата от своего среднего значения на r_{yx} его среднего квадратического отклонения.

В отличие от коэффициента регрессии b коэффициент корреляции не зависит от принятых единиц измерения признаков, а стало быть, он сравним для любых признаков.

Обычно считают связь сильной, если $r \geqslant 0{,}7$; средней — при $0{,}5 \leqslant r \leqslant 0{,}7$; слабой — при $r < 0{,}5$. Не следует гнаться за большим числом знаков при вычислении коэффициента корреляции. Во-первых, исходная информация редко имеет более трех значащих точных цифр, во-вторых, оценка тесноты связи, как правило, не требует более двух значащих цифр.

Квадрат коэффициента корреляции называется *коэффициентом детерминации:*

$$r_{yx}^2 = \frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{b\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \qquad (9.13)$$

Формула (9.13) понадобится при анализе множественной корреляции. Умножив числитель и знаменатель на $\sum_{i=1}^{n}(x_i-\bar{x})^2$, получим:

$$r_{yx}^2 = \frac{b^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

Поскольку $b(x_i-\bar{x}) = \hat{y}_i-\bar{y}$, имеем:

$$r_{yx}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \qquad (9.14)$$

Это выражение соответствует выражению η^2 (см. формулу (9.2)). Тождество коэффициента детерминации и квадрата корреляционного отношения служит основанием для интерпретации величины r_{yx}^2 как доли общей дисперсии результативного признака y, которая объясняется вариацией признака-фактора x (и связью между вариацией обоих признаков). Собственно говоря, основным показателем тесноты связи и следовало бы считать коэффициент детерминации (для линейной формулы связи) или квадрат корреляционного отношения. Но исторически раньше был введен коэффициент корреляции, который долгое время рассматривался как основной показатель.

Аналогично разным «рабочим» формулам для вычисления коэффициента регрессии можно на основе формулы (9.10) получить разные «рабочие» формулы коэффициента корреляции.

1. Разделив числитель и знаменатель формулы (9.11) на n, получим:

$$r_{yx} = \frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\sigma_y} \qquad (9.15)$$

Эта формула корреспондирует с формулой (9.8) для коэффициента регрессии.

2. Средние квадратические отклонения можно выразить через средние величины признака:

$$\sigma_x = \sqrt{\overline{x^2}-(\bar{x})^2}; \qquad \sigma_y = \sqrt{\overline{y^2}-(\bar{y})^2}.$$

Подставив эти выражения в (9.15), получим:

$$r_{yx} = \frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sqrt{[\overline{x^2}-(\bar{x})^2][\overline{y^2}-(\bar{y})^2]}}. \qquad (9.16)$$

Формула (9.16) удобнее для расчетов, если средние величины признаков и средние квадраты индивидуальных величин вычислены ранее. Смысл же коэффициента корреляции раскрывается исходной формулой (9.11). В преобразованных формулах этот смысл не столь ясен.

Пример. Рассмотрим анализ корреляционной парной линейной связи по данным 16 сельскохозяйственных предприятий о затратах на 1 корову и надое молока на 1 корову (табл. 9.1).

Таблица 9.1

Корреляция между затратами на 1 корову и надоем молока в среднем от 1 коровы

Номер единиц совокупности	Затраты на одну корову, руб./голов, x_i	Надой от одной коровы, ц, y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})\times(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	Расчетные значения надоя, ц, y_i
1	1602	34,2	−3	−1,0	+3,0	9	1,00	35,1
2	1199	19,6	−406	−15,6	+6333,6	164836	243,36	21,1
3	1321	27,3	−283	−7,9	+2235,7	80089	62,41	25,3
4	1678	32,5	+73	−2,7	−197,1	5329	7,29	37,7
5	1600	33,2	−5	−2,0	+10,0	25	4,00	35,0
6	1355	31,8	−250	−3,4	+850,0	62500	11,56	26,5
7	1413	30,7	−192	−4,5	+864,0	36864	20,25	28,5
8	1490	32,6	−115	−2,6	+299,0	13225	6,76	31,2
9	1616	26,7	+11	−8,5	−93,5	121	72,25	35,6
10	1693	42,4	+88	+7,2	+633,6	7744	51,84	38,2
11	1665	37,9	+60	+2,7	+162,0	3600	7,29	37,3
12	1666	36,6	+61	+1,4	+85,4	3721	1,96	37,3
13	1628	38,0	+23	+2,8	+64,4	529	7,84	36,0
14	1604	32,7	−1	−2,5	+2,5	1	6,25	35,2
15	2077	51,7	+472	+16,5	+7788,0	222784	272,25	51,6
16	2071	55,3	+466	+20,1	+9366,6	217156	404,01	51,4
Σ	25678	563,2	—	—	+28473,7	818533	1180,32	563,0

Средние значения признаков: $\bar{x}$ = 1605 руб./гол.; $\bar{y}$ = = 35,2 ц/голов.

Сопоставляя знаки отклонений признаков x и y от средних величин, видим явное преобладание совпадающих по знакам пар отклонений: их 14, и только 2 пары, несовпадающих знаков.

Немецкий психиатр Г. Т. Фехнер (1801—1887) предложил меру тесноты связи в виде отношения разности числа совпадающих и несовпадающих знаков пар отклонений к сумме этих чисел:

$$K_{\text{Фехнера}} = \frac{C - H}{C + H} = \frac{14 - 2}{14 + 2} = 0{,}75.$$

Конечно, коэффициент Фехнера — очень грубый показатель тесноты связи, не учитывающий величину отклонений признаков от средних значений, но он может служить некоторым ориентиром в оценке интенсивности связи. В данном случае значение коэффициента указывает на тесную связь признаков.

Вычислим на основе итоговой строки табл. 9.1 параметр уравнения парной линейной корреляции — коэффициент регрессии:

$$b = \frac{+28473{,}7}{818533} = +0{,}0347.$$

Он означает, что в среднем по изучаемой совокупности отклонение затрат на 1 корову от средней величины на 1 руб. приводило к отклонению с тем же знаком среднего надоя молока на 0,0347 ц, т.е. на 3,47 кг на корову. При нестрогой интерпретации говорят: «С увеличением затрат на одну корову на 1 руб. в среднем надой молока возрастал на 3,47 кг». Поскольку и до начала резкой инфляции стоимость 3,47 кг молока значительно превосходила рубль, увеличение затрат на 1 корову было экономически целесообразным.

Свободный член уравнения регрессии вычислим по формуле (9.6):

$$a = 35{,}2 - 0{,}0347 \cdot 1605 = -20{,}49.$$

Уравнение регрессии в целом имеет вид:

$$\tilde{y} = 0{,}0347x - 20{,}49.$$

Отрицательная величина свободного члена уравнения означает, что область существования признака y не включает нулевого значения признака x и близких к нулю значений. Можно рассчитать минимально возможную величину фактора x, при которой обеспечивается наименьшее значение признака y (разумеется, положительное).

$$x_{\min} = a : b = 20{,}49 : 0{,}0347 = 590{,}5 \text{ руб./головы}$$

— это наименьшая сумма затрат на 1 корову, при которых корова способна давать молоко. Если же область существования результативного признака включает нулевое значение признака-фактора, то свободный член является положительным и означает среднее значение результативного признака при отсутствии данного фактора, например среднюю урожайность картофеля при отсутствии органических удобрений.

Графическое изображение корреляционной связи по данным табл. 9.1 приведено на рис. 9.3.

Коэффициент корреляции, рассчитанный на основе табл. 9.1, равен:

$$r_{yx} = \frac{+28473{,}7}{\sqrt{818533 \cdot 1180{,}32}} = +0{,}916.$$

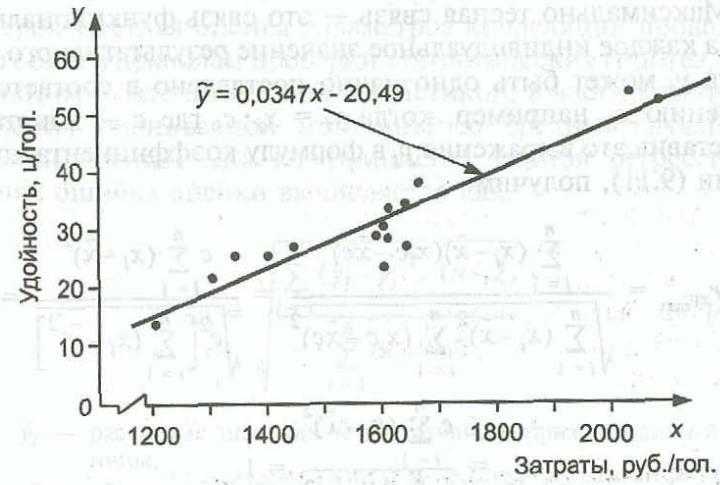


Рис. 9.3. Зависимость удойности от затрат на содержание коров

Полученное значение гораздо больше коэффициента Фехнера. Квадрат коэффициента корреляции, т.е. коэффициент детерминации, составил 0,839, или 83,9%. Вариация надоев молока на 1 корову связана с вариацией затрат в хозяйствах, произведенных в среднем на 1 корову.

Для интерпретации коэффициента корреляции необходимо знать область его существования: $0 \leqslant |r| \leqslant 1$. Как ясно из формулы (9.11), минимальное, именно нулевое, значение коэффициента корреляции может быть достигнуто, если положительные и отрицательные произведения отклонений признаков от их средних величин в числителе полностью уравновесят друг друга. Это свидетельствовало бы о полном отсутствии связи, но вероятность такого абсолютно точного взаимопогашения крайне мала для любой реальной, не бесконечно большой совокупности. Поэтому и при отсутствии реальной связи коэффициент корреляции на практике не равен нулю. Например, коэффициент корреляции между надоем молока от коров и числом букв в названии предприятия в совокупности хозяйств, указанных в табл. 9.1, равен +0,216. Как отделить реальные, надежно установленные связи от таких случайных, незначимых величин коэффициента корреляции, рассматривается в подразд. 9.5.

Максимально тесная связь — это связь функциональная, когда каждое индивидуальное значение результативного признака y_i может быть однозначно поставлено в соответствие значению x_i, например, когда $y_i = x_i \cdot c$, где c — константа. Подставив это выражение y_i в формулу коэффициента корреляции (9.11), получим:

$$r_{xy_{\max}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_ic-\bar{x}c)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(x_ic-\bar{x}c)^2}} = \frac{c\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sqrt{c^2\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}} =$$

$$= \frac{c\sum_{i=1}^{n}(x_i-\bar{x})^2}{c\sum_{i=1}^{n}(x_i-\bar{x})^2} = 1.$$

Если связь обратная и $y_i = -cx_i$, то коэффициент корреляции будет равен −1. Чем ближе коэффициент корреляции к 1, тем ближе связь к функциональной. Полученное в примере значение +0,916 свидетельствует об очень тесной связи надоев молока с затратами в расчете на 1 корову. Об этом же говорит и рис. 9.3, где реальные значения для отдельных хозяйств (точки корреляционного поля) близко расположены к линии регрессии, выражающей среднюю закономерность связи.

9.5. Статистическая оценка надежности параметров парной регрессии и корреляции

Показатели корреляционной связи, вычисленные по ограниченной совокупности (по выборке), являются лишь оценками той или иной статистической закономерности, поскольку в любом параметре сохраняется элемент не полностью погасившейся случайности, присущей индивидуальным значениям признаков. Поэтому необходима статистическая оценка степени точности и надежности параметров корреляции. Под надежностью здесь понимается вероятность того, что значение проверяемого параметра не равно нулю, не включает в себя величины противоположных знаков.

Вероятностная оценка параметров корреляции проводится по общим правилам проверки статистических гипотез, разработанным математической статистикой, в частности путем сравнения оцениваемой величины со средней случайной ошибкой оценки. Для коэффициента парной регрессии b средняя ошибка оценки вычисляется как:

$$m_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 : (n-2)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \qquad (9.17)$$

где $\hat{y}_i$ — расчетные значения результативного признака для i-й единицы;

$n-2$ — число степеней свободы (теряются 2 степени свободы, поскольку линейная парная регрессия имеет два параметра).

Числитель подкоренного выражения есть остаточная дисперсия результативного признака.

В примере по данным табл. 9.1 средняя ошибка оценки коэффициента регрессии

$$m_b = \sqrt{\frac{195{,}4:14}{818533}} = 0{,}00413.$$

Зная среднюю ошибку оценки коэффициента регрессии, можно вычислить вероятность того, что нулевое значение коэффициента входит в интервал возможных с учетом ошибки значений. С этой целью находится отношение коэффициента к его средней ошибке, т.е. t-критерий Стьюдента:

$$t = \frac{b}{m_b} = \frac{0{,}0347}{0{,}00413} = 8{,}4.$$

Табличное значение t-критерия Стьюдента при $16 - 2$ степенях свободы и уровне значимости 0,01 составляет 2,98 (табл. П.2 приложения). Полученное значение критерия намного больше, следовательно, вероятность нулевого значения коэффициента регрессии менее 0,01. Гипотезу о несущественности этого коэффициента можно отклонить: данные табл. 9.1 надежно говорят о влиянии вариации затрат на 1 корову на вариацию надоя молока от коров. Расчет t-критерия Стьюдента для коэффициента регрессии входит в программы ПЭВМ для корреляционного анализа, например «Mikrostat», «Statgraphics» и др.

Надежность установления связи можно проверить испытанием нулевой гипотезы относительно коэффициента детерминации или корреляции: $H_0 : \rho_2 = 0$. Выше показано (см. формулы (9.2) и (9.13)), что коэффициент детерминации есть доля вариации результативного признака, объясняемая его связью с вариацией одного или нескольких факторов. Из гл. 7 известна формула средней ошибки репрезентативности оценки доли по случайной выборке (формула (7.14); табл. 7.2):

$$S_p = \sqrt{\frac{p(1-p)}{n}}.$$

Применительно к коэффициенту детерминации r^2 эта формула, учитывая потерю степеней свободы вариации при парной связи, примет вид:

$$m_{r^2} = \sqrt{\frac{r^2(1-r^2)}{n-2}}. \qquad (9.18)$$

Это формула средней квадратической ошибки коэффициента детерминации. Проверим значимость заведомо несущественной корреляции надоя молока на 1 корову с числом букв в названии предприятия:

$$m_{r^2} = \sqrt{\frac{0{,}216^2(1-0{,}216^2)}{16-2}} = 0,\ 05636.$$

Критерий Стьюдента $t = \dfrac{r^2}{m_{r^2}} = \dfrac{0{,}04666}{0{,}05636} = 0{,}828.$

Величина t-критерия слишком мала, так что нулевая гипотеза не отклоняется. Измеряемая связь статистически незначима, что, собственно, было заведомо понятно.

Полученное значение t намного ниже его критического значения, которое для уровня значимости 0,1 составляет $t = 1{,}76$. Следовательно, вероятность того, что нулевое значение коэффициента входит в возможный интервал его оценок, значительно больше 0,1, и нулевая гипотеза не может быть отброшена. Конечно, анекдотический характер фактора «число букв» позволяет сделать закономерный вывод об отсутствии связи. Если же проверяемый фактор на самом деле мог влиять на результативный признак, то вывод следует формулировать не в терминах отсутствия связи, а в том, что по изучаемой информации связь надежно не установлена.

От формулы средней ошибки репрезентативности коэффициента детерминации легко перейти к формуле средней ошибки для коэффициента корреляции. Ведь для того и другого коэффициента вероятность нулевой гипотезы одна и та же. Следовательно, одинаковы и значения критерия Стьюдента:

$$t_r = t_{r^2} = \frac{r^2}{m_{r^2}} = \frac{r^2}{\sqrt{\dfrac{r^2(1-r^2)}{n-2}}} = \frac{r^2\sqrt{n-2}}{\sqrt{r^2(1-r^2)}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (9.19)$$

Поскольку $t_r = \dfrac{r^2}{m_r}$, получим формулу средней ошибки репрезентативности для коэффициента парной корреляции:

$$m_r = r : t_r = \frac{r\sqrt{1-r^2}}{r\sqrt{n-2}} = \sqrt{\frac{1-r^2}{n-2}}. \quad (9.20)$$

Если коэффициент корреляции близок к единице, то распределение его оценок отличается от нормального или распределения Стьюдента, так как его предельное значение равно единице. В таких случаях Р. Фишер предложил для оценки надежности коэффициента преобразовывать его величину в форму, не имеющую подобного ограничения:

$$z = 0{,}5 \ln \frac{1 + r_{xy}}{1 - r_{xy}}. \quad (9.21)$$

Средняя ошибка величины z определяется по формуле

$$m_z = \frac{1}{\sqrt{n-3}}. \quad (9.22)$$

Величину z можно взять из табл. П.6 приложения.

Проверим этим способом надежность коэффициента корреляции надоя молока с затратами на 1 корову:

$$z = 0{,}5 \ln \frac{1 + 0{,}916}{1 - 0{,}916} = 1{,}564;$$

$$m_z = \frac{1}{\sqrt{16-3}} = 0{,}2774; \; t = \frac{1{,}564}{0{,}2774} = 5{,}64.$$

Значение критерия Стьюдента намного больше его критического значения для уровня значимости 0,01. Следовательно, коэффициент корреляции с очень большой вероятностью больше нуля; связь установлена надежно. Для оценки надежности коэффициента корреляции можно воспользоваться таблицей критических значений для заданных уровней значимости (0,05 или 0,01) и числа степеней свободы (табл. П.5 приложения).

Например, по выборке объемом 32 единицы получен парный коэффициент корреляции 0,319. Число степеней свободы для него равно 30, поскольку в расчете r участвуют две величины, значения которых закреплены, — $\bar{x}$ и $\bar{y}$. За счет этого мы теряем две степени свободы: 32 − 2. Поскольку критическое значени r для 30 степеней свободы равно (при уровне значимости 0,05) 0,3494, то полученное значение ниже критического по модулю. Соответственно гипотеза о связи признаков надежно не доказана. Неверен будет вывод и об отсутствии связи — он также надежно не доказан. Из табл. П.5 приложения видно, что при малой выборке надежно можно установить только тесные связи, а при большой численности совокупности, например 102 единицы, надежно измеряются и слабые связи. Этот вывод важен для практической работы по корреляционному анализу.

Можно рассчитать доверительный интервал оценки коэффициента корреляции с заданной вероятностью, скажем 0,95. При этих условиях и 13 степенях свободы вариации значение t-критерия Стьюдента равно 2,16. Тогда доверительный интервал для z составит: $1{,}564 \pm 2{,}16 \cdot 0{,}2774$, т.е. от 0,965 до 2,163. Подставив эти граничные значения z в формулу (9.21), получаем границы интервала значений коэффициента корреляции: от 0,747 до 0,974. Как видим, с большой вероятностью связь на самом деле является весьма тесной, коэффициент корреляции не ниже 0,7.

9.6. Применение линейного уравнения парной регрессии

Прежде чем обсуждать вопросы использования уравнений парной регрессии, напомним, что парный корреляционный анализ не дает чистых мер влияния только одного изучаемого фактора. Если факторы взаимосвязаны, то парная связь измеряет влияние данного фактора и часть влияния прочих факторов, связанных с ним. И все же при тесной связи уравнение регрессии может стать полезным орудием анализа экономических, технологических, социальных или природных процессов.

Сравнивая фактические уровни надоя в табл. 9.1 с расчетными, т.е. такими, которые были бы получены при фактических затратах средств на 1 корову и средней по совокупности эффективности, измеряемой коэффициентом регрессии, можно найти отклонения $\tilde{y}_i - y_i$. Они показывают, насколько больше или меньше молока получило хозяйство от коров в условиях фактической эффективности использования средств, чем при средней по совокупности эффективности использо-

вания средств. Так, в хозяйстве 6 получено от 1 коровы в среднем 31,8 ц молока, хотя при низком уровне затрат 1355 руб. на 1 корову и средней эффективности затрат было бы получено только по 26,5 ц молока. Фактический надой составил 120% к расчетному. Наоборот, хозяйство 9 получало по 26,7 ц вместо расчетных 35,6 ц. Следовательно, эффективность использования средств на производство молока в этом хозяйстве (1616 руб. на 1 корову) составила только: 26,7 : 35,7 − 75% от средней по совокупности.

Оценка хозяйственной деятельности по отклонениям от расчетных значений показателей на основе уравнения регрессии (тем более на основе многофакторных регрессионных моделей) гораздо более оправданна и содержательна, чем оценка результатов производства по отклонениям от среднего значения результативного признака в совокупности, без учета факторов — характеристик возможностей и природных условий предприятия.

Уравнение регрессии применимо и для прогнозирования возможных ожидаемых значений результативного признака. При этом следует учесть, что перенос (экстраполяция) закономерности связи, измеренной в варьирующей совокупности, в статике на динамику не является, строго говоря, корректным и требует проверки условий допустимости такого решения, которое выходит за рамки статистики и может быть сделано только специалистом, хорошо знающим объект (систему) и возможности его развития.

Ограничением прогнозирования на основе регрессионного уравнения, тем более парного, служит условие стабильности или по крайней мере малой изменчивости других факторов и условий изучаемого процесса, не связанных с ними. Если резко изменится «внешняя среда» протекающего процесса, прежнее уравнение регрессии результативного признака потеряет свое значение. В засушливый год доза удобрений может не оказать влияния на урожайность сельскохозяйственной культуры, так как последнюю лимитирует недостаточная влагообеспеченность.

Прогнозируемое значение результативного показателя получается при подстановке в уравнение регрессии ожидаемой величины факторного признака. Так, если подставить в уравнение $\tilde{y} = 0{,}0347x - 20{,}49$ расход средств на одну корову, равный 2200 руб., то получим ожидаемый надой молока от коровы, равный 55,85 ц. При таком прогнозировании следует соблюдать еще одно ограничение: нельзя подставлять значения факторного признака, значительно отличающиеся от входящих, в базисную информацию, по которой вычислено уравнение регрессии. При качественно иных уровнях фактора, если они возможны в принципе, параметры уравнения были бы другими. Можно рекомендовать при определении значений факторов не выходить за пределы 1/3 размаха вариации как минимального, так и максимального значения признака-фактора, имевшегося в исходной информации.

Прогноз, полученный подстановкой в уравнение регрессии ожидаемого значения фактора, называют *точечным прогнозом*. Вероятность точной реализации такого прогноза крайне мала. Необходимо сопроводить его значением средней ошибки прогноза, или *доверительным интервалом прогноза*, с достаточно большой вероятностью. Средняя ошибка положения линии регрессии в генеральной совокупности при значении факторного признака, равном x_k, вычисляется следующим образом:

$$m_{\hat{y}(x_k)} = s_{y_{ост}} \cdot \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}, \qquad (9.23)$$

где $m_{\hat{y}(x_k)}$ — средняя ошибка положения линии регрессии в генеральной совокупности при $x = x_k$;

$s_{y_{ост}}$ — оценка среднего квадратического отклонения результативного признака от линии регрессии в генеральной совокупности с учетом степеней свободы вариации;

x_k — ожидаемое значение фактора;

n — объем выборки.

По данным табл. 9.1 находим $s_{y_{ост}}$:

$$s_{y_{ост}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{195{,}4}{16-2}} = 3{,}736 \text{ ц на 1 корову.}$$

При $x_k = 2200$ руб. на 1 голову имеем:

$$m_{\hat{y}(x_k)} = 3{,}736 \sqrt{\frac{1}{16} + \frac{(2200 - 1605)^2}{818533}} = 2{,}629 \text{ ц на 1 корову.}$$

Для вычисления доверительных границ прогноза линии регрессии нужно умножить ее среднюю ошибку на t-критерий Стьюдента. При 14 степенях свободы и доверительной вероятности 0,95 ($\alpha = 0{,}05$) значение t-критерия равно 2,14. Получаем доверительные границы: $55{,}85 \pm 2{,}629 \cdot 2{,}14$, или от 50,22 до 61,48 ц от 1 коровы. Интервал довольно широкий. Значительная неопределенность прогноза линии регрессии связана с малым объемом выборки. При объеме совокупности, равном 400, и той же вариации надоев ошибка прогноза была бы в 5 раз меньше и доверительный интервал был бы у́же.

Средняя ошибка прогноза для индивидуального значения по правилу дисперсии суммы независимых переменных образуется из ошибки прогноза положения линии регрессии и среднего квадратического отклонения индивидуальных значений от линии регрессии (остаточной вариации), т.е.

$$m_{y(x_k)} = \sqrt{m^2_{\tilde{y}(x_k)} + s^2_{y_{ост}}}. \qquad (9.24)$$

В нашем примере имеем:

$$m_{y(x_k)} = \sqrt{2{,}629^2 + 3{,}736^2} = 4{,}568 \text{ ц на 1 корову.}$$

Доверительные границы прогноза индивидуальных значений надоя молока на 1 корову при расходе 2200 руб. на 1 голову составляют с вероятностью нахождения внутри границ, равной 0,95:

$$55{,}85 \pm 4{,}568 \cdot 2{,}14, \text{ или от 46,07 до 65,63 ц.}$$

Главным источником ошибки (неопределенности) прогноза индивидуальных значений является не столько неопределенность прогноза линии регрессии, сколько значительная вариация надоев за счет других факторов, кроме входящих в уравнение регрессии.

9.7. Вычисление параметров парной линейной регрессии на основе аналитической группировки

В гл. 6 рассмотрены аналитические группировки, позволяющие установить наличие, вид и форму связи признаков. Но группировка не дает меры тесноты связи и уравнение

Таблица 9.2

Расчет по аналитической группировке

Группа предприятий по оборачиваемости в днях	Число предприятий f_j	Среднее число дней x_j'	Средняя прибыль, млн руб. $\bar{y}_j$	$\bar{y}_j - \bar{y}$	$(\bar{y}_j - \bar{y})^2 f_j$	$x' - \bar{x}$	$(x_j' - \bar{x}) \times (\bar{y}_j - \bar{y}) f_j$	$(x_j' - \bar{x})^2 f_j$	$\tilde{y}_{x_j}$	$(\tilde{y}_{x_j} - \bar{y})^2 f_j$
А	1	2	3	4	5	6	7	8	9	10
40—50	6	45	14,57	2,80	47,04	−18	−302,4	1944	15,06	64,94
51—70	8	60	12,95	1,18	11,14	−3	−28,0	72	12,36	2,78
71—101	86	6	7,40	−4,37	114,58	+23	−603,0	3174	7,69	99,88
Итого	20	63	11,77	—	172,76	—	−933,4	5190	11,77	167,60

регрессии. Теперь, пользуясь методикой корреляционно-регрессионного анализа, можно дополнить аналитическую группировку вычислением этих мер связи.

Возьмем в качестве примера группировку и рассчитаем необходимые показатели (табл. 9.2).

Коэффициент линейной регрессии

$$b = \frac{\sum_{j=1}^{3} (x_j' - \bar{x})(\bar{y}_j - \bar{y}) f_j}{\sum_{j=1}^{3} (\bar{x}_j - \bar{x})^2 f_j} = \frac{-933{,}4}{5190} = -0{,}18,$$

Свободный член уравнения регрессии

$$a = \bar{y} - b\bar{x} = 11{,}77 - (-0{,}18 \cdot 63) = 23{,}15.$$

Итак, имеем уравнение связи: $\tilde{y} = 23{,}15 - 0{,}18x$. Вычислим по этому уравнению расчетные значения прибыли $\tilde{y}_j$ для каждой группы. Подставив в уравнение середины интервалов групп x_j', запишем $\tilde{y}_j$ в гр. 9 табл. 9.2. Вариация расчетных значений прибыли связана с влиянием оборачиваемости x. Найдем сумму квадратов отклонений прибыли за счет вариации оборачиваемости — факторную вариацию (графа 10 табл.

9.2). Для расчета общей вариации результативного признака была вычислена сумма квадратов отклонений по индивидуальным данным:

$$\sum_{1}^{20}(y_i - \bar{y})^2.$$

Эта сумма квадратов — общая вариация объема прибыли — равна 222,4. Теперь можем построить меры тесноты связи:

- *теоретическое корреляционное отношение*

$$\eta_т^2 = \frac{\sum_{j=1}^{3}(\bar{y}_j - \bar{y})^2 f_j}{\sum_{1}^{20}(y_j - \bar{y})^2} = \frac{167,6}{222,4} = 0,754;\ \eta_т = \sqrt{0,754} = 0,868;$$

- *эмпирическое корреляционное отношение*

$$\eta_э = \sqrt{\frac{172,76}{222,4}} = \sqrt{0,777} - 0,881.$$

Оба квадрата корреляционных отношений соответствуют по содержанию рассмотренному коэффициенту детерминации (9.1) и (9.2) и интерпретируются как показатели доли вариации результативного признака, объясняемой за счет вариации группировочного, факторного признака (и, конечно, связанных с ним прочих факторов). В данном примере связь является тесной. Различие в том, что в эмпирическом корреляционном отношении связь признаков не абстрагирована от случайных влияний прочих факторов на вариацию y, не связанных с вариацией x.

Наиболее рациональным приемом анализа и расчета параметров корреляционной связи с помощью группировки является построение так называемой *корреляционной решетки* (табл. 9.3). Это таблица, в которой изучаемая совокупность сгруппирована одновременно по обоим признакам, связь между которыми изучается (двумерное распределение). Число групп по признакам может быть как равным, так и неравным. Если наибольшие значения частот каждой строки и каждого столбца располагаются на первой диагонали (в рассматриваемой таблице эти цифры выделены), связь является прямой и близ-

Таблица 9.3

Корреляция между возрастом женихов и невест, вступивших в брак

Возраст женихов, лет x	Середины интервалов x_i'	Возраст невест, лет, y					Итого f_j	$x_i' - \bar{x}$	$(x_i' - \bar{x})^2$	$(x_i' - \bar{x})^2 f_i$
		до 25 $y_1' = 21$	25—34 $y_2' = 30$	35—44 $y_3' = 40$	45—54 $y_4' = 50$	55 и старше $y_5' = 62$				
До 25	22	**18 212**	1914	147	8	**4**	20 285	−9,2	84,64	1716922,4
25—34	30	5574	**6677**	1112	**85**	18	13 466	−1,2	1,44	19391,04
35—44	40	498	2171	**2595**	419	43	5 726	+8,8	77,44	443421,44
45—54	50	98	**368**	1177	**1280**	308	3 231	+18,8	353,44	1141964,6
55 и старше	63	**19**	75	271	840	**1701**	2 906	+31,8	1011,24	2938663,4
Итого f_j		24 401	11 205	5302	2632	2074	45 614			6260362,8
$y_j' - \bar{y}$		−9	+1	+11	+21	+33				
$(y_j' - \bar{y})^2$		81	1	121	441	1089				
$(y_j' - \bar{y})^2 f_j$		1 976 481	11 205	641 542	1 160 712	2 258 586	6 048 526			5196031,6

Примечание. Средний возраст составил для женихов: $\bar{x} = 31,2$ года, для невест: $\bar{y} = 29,0$ года.

кой к линейной; если наибольшие значения частот располагаются вдоль второй диагонали (в рассматриваемой таблице эти цифры также выделены), связь обратная, линейная. Если частоты во всех клетках таблицы примерно равны, связи нет; если наибольшие значения расположены по дуге, связь криволинейная. В табл. 9.3, кроме частот, приведены строки и графы для расчета необходимых сумм при вычислении параметров корреляционной связи.

В данной таблице наибольшие частоты в строках и графах расположены на первой диагонали, что говорит в соответствии с логикой о прямой линейной связи возрастов женихов и невест. Связь эта далеко не полная; как видим, «любви все возрасты покорны», все клетки таблицы заполнены, значит, существуют браки между лицами любых возрастов.

Как средние величины признаков, так и все суммы, входящие в расчет параметров корреляции, при группировке взвешиваются на соответствующие частоты, поэтому формулы (9.9) и (9.11) приобретают следующий вид:

$$b = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}(x_i'-\bar{x})(y_j'-\bar{y})f_{ij}}{\sum_{i=1}^{k_1}(x_i'-\bar{x})^2 f_i}; \qquad (9.25)$$

$$r_{x_1y} = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}(x_i'-\bar{x})(y_j'-\bar{y})f_{ij}}{\sqrt{\sum_{i=1}^{k_1}(x_i'-\bar{x})^2 f_i \cdot \sum_{j=1}^{k_2}(y_j'-\bar{y})f_j}}, \qquad (9.26)$$

где x_i', y_j' — середины интервалов i-й категории x и j-й категории y;

f_i — частота i-го значения x;

f_j — частота j-го значения y;

f_{ij} — частота совместного появления i-го значения x и j-го значения y (числа в клетках корреляционной решетки).

Взвешенные суммы квадратов отклонений подсчитаны и приведены в последней графе и в последней строке табл. 9.3. Для вычисления числителя в (9.25) и (9.26) необходимо умножить отклонения по обоим признакам (с учетом их знаков)

на частоты совместного распределения и сложить все 25 произведений:

$$(-9)\cdot(-9{,}2)\cdot 18212 + 1\cdot(-9{,}2)\cdot 1914 + 11\cdot(-9{,}2)\cdot\ 147 + \\ + \ldots 21\ \cdot\ 18{,}8\ \cdot\ 1280 + \ldots + 33\cdot 31{,}8\cdot 1701 = 5196031{,}6.$$

Это число записано в правом нижнем углу табл. 9.3. Рассчитаем параметры уравнения регрессии. Согласно (9.25)

$$b = \frac{5196031{,}6}{6260362{,}8} = 0{,}830.$$

Это означает, что в среднем с увеличением возраста женихов на 1 год возраст невест увеличивался на 0,83 года. Свободный член уравнения согласно (9.6)

$$a = 29{,}0 - 0{,}83\cdot 31{,}2 = 3{,}1.$$

Уравнение имеет вид:

$$\hat{y} = 3{,}1 + 0{,}83\cdot x.$$

Поскольку оба признака равноправны, можно получить уравнение зависимости среднего возраста жениха от возраста невесты. Поменяв местами x и y, получаем:

$$b = \frac{5196031{,}6}{6048526} = 0{,}859;\ a = 31{,}2 - 0{,}859\cdot 29 = 6{,}3;\ \ \hat{x} = 6{,}3 + 0{,}859y,$$

т.е. в среднем с увеличением возраста невесты на 1 год возраст жениха возрастал на 0,86 года.

Коэффициент корреляции согласно формуле (9.26) составляет:

$$r_{xy} = \frac{5196031{,}6}{\sqrt{6260362{,}8\cdot 6048526}} = 0{,}844.$$

Коэффициент детерминации $r^2 = 71{,}3\%$, т.е. вариация возраста супруга или супруги на 71% зависит от вариации возраста «второй половины». Связь весьма тесная.

Конечно, расчет параметров корреляции на основе группировки является приближенным: реальные значения признаков заменяются серединами интервалов, а при открытых интервалах — их экспертными оценками. Не учитывается неравномерность изменения частот внутри интервалов. Казалось бы, с появлением ППП этот метод должен отмереть. Од-

нако для больших совокупностей ППП имеют ограничения на объем оперативной памяти. Вдобавок корреляционные решетки очень наглядны, и специалист по расположению клеточных частот может сделать заключение о тесноте связи признаков.

9.8. Параболическая корреляция

Линейные связи являются основными. Однако встречаются и нелинейные связи, хорошо описываемые параболой, гиперболой и т.д.

Уравнение регрессии в форме параболы 2-го порядка имеет следующий вид:

$$\hat{y} = a + bz + cx^2. \qquad (9.27)$$

Если при линейной связи среднее изменение результативного признака на единицу фактора постоянно по всей области вариации фактора, то при параболической корреляции изменение признака x на единицу признака y меняется равномерно с изменением величины фактора. В результате связь может даже поменять знак на противоположный, из прямой превратится в обратную, из обратной в прямую. Такой характер связи объективно присущ многим системам. Например, с увеличением дозы удобрений урожайность сельскохозяйственных культур сначала повышается, но если превысить оптимальную величину дозы, то при дальнейшем росте дозы удобрений растения угнетаются и урожайность снижается.

Нормальные уравнения при использовании метода наименьших квадратов для нахождения параметров уравнения параболы 2-го порядка таковы:

$$na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i;$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} y_i x_i;$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} y_i x_i^2.$$

Если расчет производится не по индивидуальным данным, а на основе аналитической группировки, то система уравнений МНК приобретают следующий вид:

$$a\sum_{j=1}^{k} f_j + b\sum_{j=1}^{k} x_j' f_j + c\sum_{j=1}^{k} x_j'^2 f_j = \sum_{j=1}^{k} y_j' f_j;$$

$$a\sum_{j=1}^{k} x_j' f_j + b\sum_{j=1}^{k} x_j'^2 f_j + c\sum_{j=1}^{k} x_j'^3 f_j = \sum_{j=1}^{k} y_j' x_j' f_j;$$

$$a\sum_{j=1}^{k} x_j'^2 f_j + b\sum_{j=1}^{k} x_j'^3 f_j + c\sum_{j=1}^{k} x_j'^4 f_j = \sum_{j=1}^{k} y_j' x_j'^2 f_j.$$

Решая эту систему, получаем значения параметров a, b и c. Показателем тесноты параболической корреляции является корреляционное отношение, вычисляемое как корень квадратный из выражения (9.2).

Пример. Рассмотрим параболическую корреляционную связь на примере зависимости себестоимости молока от продуктивности коров по данным аналитической группировки сельскохозяйственных предприятий области (табл. 9.4). В этой же таблице приведены расчетные величины, входящие в уравнения МНК для параболы.

Была получена система нормальных уравнений:

$$136a + 525b + 2123{,}4c = 4585{,}1;$$

$$525a + 2123{,}4b + 9017{,}1c = 17318{,}1;$$

$$2123{,}4a + 9017{,}1b + 40199{,}3c = 68586{,}7.$$

Решив систему уравнений, получим:

$$c = 2{,}3249;\ b = -23{,}64;\ a = 88{,}68.$$

Уравнение регрессии имеет вид:

$$\hat{y} = 88{,}68 - 23{,}641x + 2{,}3249x^2. \qquad (9.28)$$

Эта парабола имеет точку минимума в фактической области вариации факторного признака. Для нахождения значения фактора, при котором достигается минимальное значение результативного признака, следует приравнять нулю первую производную по x уравнения (9.28):

$$\frac{d\hat{y}}{dx} = -23{,}641 + 2 \cdot 2{,}3249x,$$

откуда: $x = 23{,}641/4{,}6498 = 5{,}084$ т молока на 1 корову.

Таблица 9.4

Определение параметров параболической зависимости себестоимости молока от продуктивности коров

Группы предприятий по надою т, x	Число предприятий, f_j	Середина интервалов т, x'_j	Средняя себестоимость молока, руб./т, $\bar{y}_j$	$x'_j f_j$	$x'^2_j f_j$	$x'^3_j f_j$	$x'^4_j f_j$	$\bar{y}_j f_j$	$\bar{y}_j f_j x'_j$	$\bar{y}_j f_j x'^2_j$	$\tilde{y}_j$	$\tilde{y}_j - \bar{y}$	$(\tilde{y}_j - \bar{y})^2 \times f_j$
До 3,0	16	2,80	41,25	44,8	125,4	351,2	983,4	660,0	1848,0	5174,4	40,71	+7,00	784,0
3,0—3,5	42	3,25	36,67	136,5	443,6	1441,8	4685,8	1540,1	5005,5	16267,7	36,40	+2,69	303,9
3,5—4,0	31	3,75	32,50	116,2	435,9	1634,8	6130,4	1007,5	3778,1	14168,0	32,72	−0,99	30,4
4,0—4,5	18	4,25	28,89	76,5	325,1	1381,8	5872,6	520,0	2210,1	9392,9	30,20	−3,51	221,8
4,5—5,0	13	4,75	29,04	61,8	293,3	1393,2	6617,9	377,5	1793,2	8517,8	28,84	−4,87	308,3
5,0—5,5	9	5,25	29,17	47,2	248,1	1302,3	6837,2	262,5	1378,3	7236,3	28,64	−5,07	321,3
Более 5,5	7	6,00	31,07	42,0	252,0	1512,0	9072,0	217,5	1304,9	7829,6	30,53	−3,18	70,8
Итого	136	—	33,71	525,0	2123,4	9017,1	40199,3	4585,1	17318,1	68586,7	—	—	1950,5

Итак, минимальная себестоимость молока в совокупности предприятий, в условиях периода, к которому относятся данные, достигалась в среднем при надое молока на 1 корову 5084 кг. Значение фактора x при достижении минимума себестоимости можно назвать оптимальной продуктивностью коров, а саму задачу его поиска — одной из оптимизационных задач, решаемых математико-статистическим методом.

Для измерения тесноты параболической корреляционной связи находим вариацию результативного признака y, объясняемую вариацией фактора x, как сумму квадратов отклонений расчетных величин $\tilde{y}$ от средней величины $\bar{y}$, взвешенных на число предприятий. Общая сумма квадратов отклонений всех 136 значений y_i от средней величины составляет 4624,7. Таким образом, согласно формуле (9.1) корреляционное отношение

$$\eta = \sqrt{\frac{1950,5}{4624,7}} = \sqrt{0,422} = 0,650.$$

Поскольку η^2 — аналог коэффициента детерминации, можно сделать вывод, что 42,2% вариации себестоимости молока в совокупности 136 предприятий были связаны с вариацией продуктивности коров (и с факторами, варьирующими согласованно с продуктивностью, в соответствии с ранее сделанной оговоркой об интерпретации парных связей).

9.9. Гиперболическая корреляция

Уравнение регрессии в форме гиперболы имеет следующий вид:

$$\hat{y} = a + \frac{b}{x}. \qquad (9.29)$$

Если величина b положительна, то при увеличении значений факторного признака x значения результативного признака уменьшаются, причем это уменьшение все время замедляется, и при $x \to \infty$ средняя величина признака y будет равна a. Если же параметр b отрицателен, то значения результативного признака с ростом фактора возрастают, причем их рост замедляется, и в пределе при $x \to \infty$ $\tilde{y} = a$. Таким образом, гипер-

болические зависимости характерны для связей, в которых результативный признак не может варьировать неограниченно, его вариация имеет односторонний предел. Например, при освоении нового оборудования его производительность возрастает, но рост замедлится по мере приближения к конструктивно-технологическому пределу производственной мощности агрегата. Совершенствуя двигатель, можно увеличивать его КПД, но тоже не выше предела, допустимого данным видом преобразования энергии. Таков же характер связи между уровнем душевого дохода x в семье и долей семей, имеющих телевизоры, y; он приближен к пределу (100%) в наиболее обеспеченной группе семей.

Нормальные уравнения метода наименьших квадратов для гиперболы таковы:

$$\begin{cases} na + b\sum_1^n \frac{1}{x_i} = \sum_1^n y_i; \\ a\sum_1^n \frac{1}{x_i} + b\sum_1^n \frac{1}{x_i^2} = \sum_1^n \frac{y_i}{x_i}. \end{cases}$$

Легко видеть, что эти уравнения, по существу, те же, что и для линейной связи. Линеаризация гиперболического уравнения достигается заменой $1/x$ на новую переменную, которую можно обозначить z. Тогда уравнение (9.29) примет вид: $\tilde{y} = a + bz$. Это и следует сделать, вычисляя гиперболу на компьютере, если программа не предусматривает автоматического вычисления гиперболических регрессий.

Пример. Рассмотрим расчет уравнения гиперболической связи на примере анализа влияния среднесуточного прироста живой массы крупного рогатого скота на откорме на себестоимость прироста живой массы в совокупности предприятий области, занимавшихся откормом скота (табл. 9.5).

Нормальные уравнения МНК имеют вид:

$$\begin{cases} 123a + 24{,}03b = 47634; \\ 24{,}03a + 4{,}88b = 9650. \end{cases}$$

Решение системы уравнений дает значения показателей:

$$a = 24{,}44;\ b = 1857.$$

Таблица 9.5

Гиперболическая связь себестоимости прироста со скоростью прироста массы скота

Группы предприятий по среднесуточному приросту массы граммов на 1 голову x_j	Число предприятий f_j	Средняя себестоимость прироста, руб./ц $\bar{y}_j$	Середина интервалов x_j' в сотнях граммов на 1 голову	$\frac{f_j}{x_j'}$	$\frac{f_j}{(x_j')^2}$	$\bar{y}_j f_j$	$\frac{\bar{y}_j f_j}{x_j'}$	$\bar{y}_j$
334—425	22	496	3,8	5,79	1,52	10912	2872	513
425—516	37	425	4,7	7,87	1,67	15725	3346	419
516—607	28	360	5,6	5,00	0,89	10080	1800	356
607—698	27	310	6,5	4,15	0,64	8370	1288	310
698—789	9	283	7,4	1,22	0,16	2547	344	275
Итого	123	387	—	24,03	4,88	47634	9650	—

Уравнение регрессии имеет вид:

$$\tilde{y} = 24{,}44 + \frac{1857}{x},$$

где x — в сотнях граммов.

Точечный прогноз по уравнению регрессии при среднесуточном приросте массы животных, равном 900 г, уже достигнутом передовыми хозяйствами, приводит к ожидаемой средней себестоимости: $\tilde{y} = 24{,}44 + \frac{1857}{9} = 230{,}77$ руб./ц.

Корреляционное отношение для данной связи, рассчитанное по формуле (9.1), равно:

$$\eta = \sqrt{\frac{687047}{1023451}} = \sqrt{0{,}671} = 0{,}819.$$

Следовательно, 67% вариации себестоимости прироста массы скота объяснялись вариацией скорости роста массы и связанных с ней других факторов, например, чем быстрее растет масса, тем меньше расход кормов на 1 прироста массы.

9.10. Множественное уравнение регрессии

Проблемы множественного корреляционно-регрессионного анализа и моделирования обычно подробно изучаются в специальном курсе. В курсе «Общая теория статистики» рассматриваются только самые общие вопросы этой сложной проблемы и дается начальное представление о методике построения уравнения множественной регрессии и показателей связи. Рассмотрим линейную форму многофакторных связей не только как наиболее простую, но и как форму, предусмотренную пакетами прикладных программ для ПЭВМ. Если же связь отдельного фактора с результативным признаком не является линейной, то проводят линеаризацию уравнения путем замены или преобразования величины факторного признака.

Общий вид многофакторного уравнения регрессии следующий:

$$\hat{y} = a + b_1x_1 + \ldots + b_kx_k = a + \sum_{j=1}^{k} b_jx_j, \qquad (9.30)$$

где k — число факторных признаков (независимых переменных).

Для того чтобы упростить систему уравнений МНК, необходимую для вычисления параметров уравнения (9.30), обычно вводят величины отклонений индивидуальных значений всех признаков от средних величин этих признаков.

$$\Delta y_i = y_i - \bar{y}; \Delta x_{ij} = x_{ji} - \bar{x}_j.$$

Получаем систему k уравнений МНК:

$$b_1 \sum_{i=1}^{n} \Delta^2 x_{1i} + b_2 \sum_{i=1}^{n} \Delta x_{1i}\Delta x_{2i} + \ldots + b_k \sum_{i=1}^{n} \Delta x_{1i}\Delta x_{ki} = \sum_{i=1}^{n} \Delta y_i \Delta x_{1i};$$

$$b_1 \sum_{i=1}^{n} \Delta x_{1i}\Delta x_{2i} + b_2 \sum_{i=1}^{n} \Delta^2 x_{2i} + \ldots + b_k \sum_{i=1}^{n} \Delta x_{2i}\Delta x_{ki} = \sum_{i=1}^{n} \Delta y_i \Delta x_{2i};$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$b_1 \sum_{i=1}^{n} \Delta x_{1i}\Delta x_{ki} + b_2 \sum_{i=1}^{n} \Delta x_{2i}\Delta x_{ki} + \ldots + b_k \sum_{i=1}^{n} \Delta^2 x_{ki} = \sum_{i=1}^{n} \Delta y_i \Delta x_{ki}.$$

Решив эту систему, получим значения *коэффициентов условно-чистой регрессии* b_j. Свободный член уравнения вычисляется по формуле

$$a = \bar{y} - \sum_{j=1}^{k} b_j\bar{x}_j. \qquad (9.31)$$

Термин «коэффициент условно-чистой регрессии» означает, что каждая из величин b_j измеряет среднее по совокупности отклонение результативного признака от его средней величины при отклонении данного фактора x_j от своей средней величины на единицу его измерения и при условии, что все прочие факторы, входящие в уравнение регрессии, закреплены на средних значениях (не изменяются, не варьируют).

Таким образом, в отличие от коэффициента парной регрессии коэффициент условно-чистой регрессии измеряет влияние фактора, абстрагируясь от связи вариации этого фактора с вариацией остальных факторов. Если было бы возможным включить в уравнение регрессии все факторы, влияющие на вариацию результативного признака, то величины b_j можно было бы считать мерами *чистого* влияния факторов. Но так как реально невозможно включить все факторы в уравнение, то коэффициенты b_j не свободны от примеси влияния факторов, не входящих в уравнение.

Включить все факторы в уравнение регрессии невозможно по одной из трех причин или сразу по всем, так как: 1) часть факторов может быть неизвестна современной науке, познание любого процесса всегда неполное; 2) по части известных теоретически факторов нет информации либо таковая ненадежна; 3) численность изучаемой совокупности (выборки) ограниченна, что позволяет включить в уравнение регрессии ограниченное число факторов.

Коэффициенты условно-чистой регрессии b_j являются именованными числами, выраженными в разных единицах измерения, и поэтому несравнимы друг с другом. Для преобразования их в сравнимые относительные показатели применяется то же преобразование, что и для получения коэффициента парной корреляции. Полученную величину называют *стандартизованным коэффициентом регрессии или β-коэффициентом.*

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}. \quad (9.32)$$

Данный β_j-коэффициент при факторе x_j определяет степень влияния вариации фактора x_j на вариацию результативного признака y при отвлечении от сопутствующей вариации других факторов, входящих в уравнение регрессии.

Коэффициенты условно-чистой регрессии полезно выразить в виде относительных сравнимых показателей связи, *коэффициентов эластичности:*

$$e_j = b_j \frac{\bar{x}_j}{\bar{y}}. \quad (9.33)$$

Коэффициент эластичности фактора x_j говорит о том, что при отклонении величины данного фактора от средней величины на 1% и при отвлечении от сопутствующего отклонения других факторов, входящих в уравнение, результативный признак отклонится от своего среднего значения на e_j процентов от $\bar{y}$. Чаще интерпретируют и употребляют коэффициенты эластичности в терминах динамики: при увеличении фактора x_j на 1% его средней величины результативный признак увеличится на e_j процентов его средней величины.

Рассмотрим расчет и интерпретацию уравнения многофакторной регрессии на примере тех же 16 хозяйств, приведенных в табл. 9.6 (результативный признак — уровень валового дохода и три фактора, влияющих на него).

Напомним еще раз, что для получения надежных и достаточно точных показателей корреляционной связи необходима более многочисленная совокупность.

Решение проведено по программе «Microstat» для ПЭВМ. Приведем таблицы из распечатки: табл. 9.7 включает средние величины и средние квадратические отклонения всех признаков. Она содержит коэффициенты регрессии и их вероятностную оценку: 1-я графа «Var.» — переменные, т.е. факторы; 2-я «Regression coefficient» — коэффициенты условно-чистой регрессии b_j; 3-я «Std. error» — средние ошибки оценок коэффициентов регрессии; 4-я — значения t-критерия Стьюдента при 12 степенях свободы вариации; 5-я «Prob.» — вероятности отклонения нулевой гипотезы относительно коэффициентов регрессии; 6-я графа «Partial r^2» — частные коэффициенты детерминации. Содержание и методика расчета показателей в графах 3—6 рассматриваются далее в гл. 9 «Constant» — свободный член уравнения регрессии a; «Std. error of est.» — средняя квадратическая ошибка оценки результативного признака по уравнению регрессии.

Было получено уравнение множественной регрессии:

$$\hat{y} = 2{,}26x_1 - 4{,}31x_2 + 0{,}166x_3 - 240.$$

Это означает, что величина валового дохода на 1 га сельскохозяйственных угодий в среднем по совокупности возрастала на 2,26 руб. при увеличении затрат труда на 1 чел.-день/га; уменьшалась в среднем на 4,31 руб. при возрастании доли

Таблица 9.6

Уровень валового дохода и его факторы

Номер хозяйства	Валовой доход, руб./га, y	Затраты труда, чел.-дн./га, x_1	Доля пашни, %, x_2	Надой молока на 1 корову, кг, x_3
1	704	265	45,1	3422
2	293	193	35,1	1956
3	346	229	69,4	2733
4	420	193	60,2	3254
5	691	225	59,0	3323
6	679	255	63,4	3179
7	457	201	58,1	3073
8	503	208	51,8	3257
9	314	170	73,2	2669
10	803	276	59,0	4235
11	691	188	42,5	3790
12	775	232	50,5	3658
13	584	173	48,6	3801
14	504	183	51,9	3266
15	777	236	58,9	5173
16	1138	265	38,8	5526
Сумма	9679	3492	865,5	56315
Средняя	604,9	218,2	54,1	3520
s	221,9	34,6	10,6	887
v,%	36,7	15,9	19,6	25,2

Таблица 9.7

Показатели уравнения регрессии

Dependent variable: y					
Var.	Regression coefficient	Std. error	t ($d.f.$ = 12)	Prob.	Partial r^2
1	2	3	4	5	6
x_1	2,260978	0,680030	3,325	0,00606	0,4795
x_2	−4,307303	1,982283	−2,173	0,05053	0,2824
x_3	0,166091	0,027050	6,140	0,00005	0,7586
Constant −240,112905 Std. error of est. = 79,243276					

пашни в сельскохозяйственных угодьях на 1% и увеличивалась на 0,166 руб. при росте надоя молока на 1 корову на 1 кг. Отрицательная величина свободного члена вполне закономерна, и, как уже отмечено в подразд. 9.2, нулевые значения факторов в производстве невозможны, и свободный член выполняет роль доводки до функционального соотношения между средними величинами и экономического смысла не имеет:

$$a = \bar{y} - \sum_{(j)} b_j \bar{x}_j.$$

Отрицательное значение коэффициента при x_2 — сигнал о неблагополучии в экономике изучаемых хозяйств, где растениеводство убыточно, а прибыльно только животноводство. При рациональных методах ведения сельского хозяйства и нормальных ценах (равновесных или близких к ним) на продукцию всех отраслей доход должен не уменьшаться, а возрастать с увеличением пашни — наиболее плодородной части сельскохозяйственных угодий.

По данным двух предпоследних строк табл. 9.6 и табл. 9.7 рассчитаем β-коэффициенты и коэффициенты эластичности согласно формулам (9.32) и (9.33).

Как на вариацию уровня дохода, так и на его возможное изменение в динамике самое сильное влияние оказывает фактор x_3 — продуктивность коров, а самое слабое — x_2 — доля пашни. Значения β_j^2 будут использоваться в дальнейшем (табл. 9.8).

Таблица 9.8

Сравнительное влияние факторов на уровень дохода

Факторы x_j	β_j	e_j	β_j^2
x_1	0,352	0,816	0,138
x_2	−0,206	−0,385	0,042
x_3	0,664	0,966	0,441

Между β-коэффициентом и коэффициентом эластичности существует следующее соотношение:

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y};$$

в качестве σ_{x_j}, σ_y приняты их оценки s_{x_j}, s_y (табл. 9.6):

$$e_j = b_j \frac{\bar{x}_j}{\bar{y}}.$$

Разделив β_j на e_j, имеем: $\frac{\beta_j}{e_j} = b_j \frac{s_{x_j}}{s_y} : b_j \frac{\bar{x}_j}{\bar{y}} = \frac{s_{x_j}}{\bar{x}_j} : \frac{s_y}{\bar{y}} = v_{x_j} : v_y$. (9.34)

Итак, мы получили, что β-коэффициент фактора x_j относится к коэффициенту эластичности этого фактора, как коэффициент вариации фактора x к коэффициенту вариации результативного признака y. Поскольку, как видно из последней строки табл. 9.6, коэффициенты вариации всех факторов меньше коэффициента вариации результативного признака, все β-коэффициенты меньше коэффициентов эластичности.

Рассмотрим соотношение между парным и условно-чистыми коэффициентами регрессии на примере фактора x_1. Парное линейное уравнение связи y с x_1 имеет вид:

$$\hat{y} = 3{,}886x_1 - 243{,}2.$$

Условно-чистый коэффициент регрессии при x_1 составляет только 58% парного ($b_{1,23} = 2{,}26$). Остальные 42% связаны с тем, что вариации x_1 сопутствует вариация факторов x_2, x_3, которые, в свою очередь, влияют на результативный признак

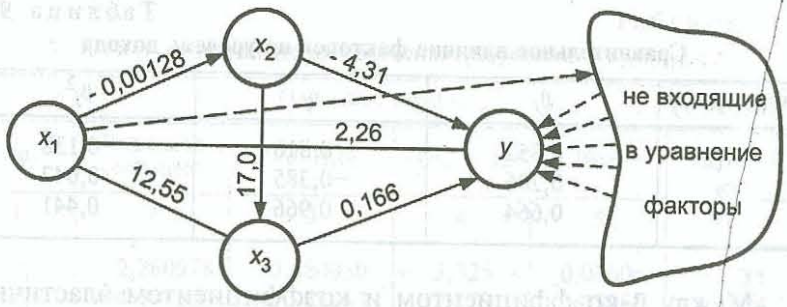


Рис. 9.2. Граф связей (по данным табл. 9.7)

y. Связи всех признаков и их коэффициенты парных регрессий представлены на рис. 9.2.

Если сложить оценки прямого и опосредованного влияния вариации x_1 на y, т.е. произведения коэффициентов парных регрессий по всем «путям» (см. рис. 9.2), получим: $2{,}26 + 12{,}55 \cdot 0{,}166 + (-0{,}00128) \cdot (-4{,}31) + (-0{,}00128) \cdot 17{,}00 \times 0{,}166 = 4{,}344$.

Эта величина даже больше парного коэффициента связи x_1 с y. Следовательно, косвенное влияние вариации x_1 через не входящие в уравнение признаки-факторы — обратное, дающее в сумме снижение оценки влияния x_1 на $-0{,}458$:

$$3{,}886 - 4{,}344 = -0{,}458.$$

9.11. Меры тесноты связей в многофакторной системе

Многофакторная система требует уже не одного, а множества показателей тесноты связей, имеющих разный смысл и применение. Основой измерения связей является матрица парных коэффициентов корреляции (табл. 9.9).

По этой матрице можно судить о тесноте связи факторов с результативным признаком и между собой. Хотя все эти показатели относятся к парным связям, все же матрицу можно использовать для предварительного отбора факторов для включения их в уравнение регрессии. Не рекомендуется включать в уравнение факторы, слабо связанные с результативными признаками, но тесно связанные с другими факто-

Таблица 9.9

Матрица парных коэффициентов корреляции (общий вид)

Признак	y	x_1	...	x_k
y	1			
x_1	r_{yx_1}	1		
.	⋮	⋮		
.	r_{yx_j}	$r_{x_1x_j}$	1 ...	
.	⋮	⋮	...	
x_k	r_{yx_k}	$r_{x_1x_k}$	$r_{x_jx_k}$	1 ...

Таблица 9.10

Матрица парных коэффициентов корреляции

Признак	y	x_1	x_2	x_3
y	1			
x_1	0,687	1		
x_2	−0,355	−0,044	1	
x_3	0,878	0,490	−0,203	1

рами. Если, например, имеем: $r_{yx_1} = 0{,}8$; $r_{yx_2} = 0{,}65$; $r_{x_1x_2} = 0{,}88$, то в регрессионное уравнение следует включить фактор x_1, а фактор x_2 не включать, так как он тесно связан с x_1 (коллинеарен с x_1) и его корреляция с y слабее, чем корреляция фактора x_1. Совершенно недопустимо включать в анализ факторы, функционально связанные друг с другом, т.е. с коэффициентом корреляции, равным (или близким) 1. Включение таких пар признаков приводит к вырожденной матрице коэффициентов корреляции и неопределенности решения. В этом случае решение задачи на ПЭВМ прекращается.

Матрица парных коэффициентов для нашего примера (табл. 9.10) говорит об отсутствии коллинеарных (т.е. линейно связанных) факторов, что позволяет включить все эти факторы в уравнение регрессии.

На основе этой матрицы вычисляется наиболее общий показатель тесноты связи всех входящих в уравнение регрессии факторов с результативным признаком — коэффициент мно-

жественной детерминации $R^2_{y,x_1\ldots x_k}$ как частное от деления определителя матрицы Δ^* на определитель матрицы Δ, где

$$\Delta^* = \begin{vmatrix} r_{yx_1} & r_{yx_2} & \ldots & r_{yx_k} & 0 \\ 1 & r_{x_1x_2} & \ldots & r_{x_1x_k} & r_{yx_1} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_k} & r_{yx_2} \\ \vdots & \vdots & & \vdots & \vdots \\ r_{x_1x_k} & r_{x_2x_k} & \ldots & 1 & r_{yx_k} \end{vmatrix};$$

$$\Delta = \begin{vmatrix} 1 & r_{x_1x_2} & \ldots & r_{x_1x_k} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_k} \\ \vdots & \vdots & & \vdots \\ r_{x_1x_k} & r_{x_2x_k} & \ldots & 1 \end{vmatrix}.$$

Этим способом можно определить величину R^2, не вычисляя расчетных значений результативного признака $\hat{y}_i$ для всех единиц совокупности. Если полученная величина R^2 не удовлетворяет исследователя, то можно прекратить дальнейшие вычисления и не рассчитывать $\hat{y}_i$ (это имеет значение, если совокупность состоит из сотен и тысяч единиц).

Принципиальное содержание множественного коэффициента детерминации, как и парного, раскрывается формулой (9.2). Это отношение части вариации результативного признака, объясняемой за счет вариации входящих в уравнение факторов, к общей вариации результативного признака за счет всех факторов. Здесь под вариацией понимается сумма квадратов отклонений индивидуальных расчетных по уравнению величин от средней («объясненная вариация») и первичных индивидуальных величин от средней («общая вариация»).

В нашем примере значение сумм квадратов отклонений и коэффициенты детерминации и корреляции приведены по распечатке программы «Microstat» в табл. 9.11.

Верхняя строка: корректированный R-квадрат = 0,872390; вторая строка: R-квадрат = 0,897912; третья строка: множественный R = 0,947582. Ниже приводится таблица дисперсионного анализа, в которой указываются источники вариации: *объясненная* сумма квадратов отклонений значений, рассчитанных

Таблица 9.11

Показатели множественной корреляционной связи

Adjusted R squared = 0.872390
R squared = 0.897912
Multiple R = 0.947582

Analysis of variance table

	Sum of squares	*d.f.*	Mean square	F-Ratio	Prob
Regression	662772.975894	3	220924.325298	35.182	3.171E−06
Residual	75353.961606	12	6279.496801		
Total	738126.937500	15			

по уравнению регрессии, от среднего значения: $D_{\text{объясн}} = \Sigma(\hat{y}_i - \bar{y})^2 = 662772{,}98$ при числе степеней свободы, равном числу объясняющих переменных $d.f. = 3$; *остаточная* — отклонения фактических значений от расчетных: $D_{\text{ост}} = \Sigma(y_i - \hat{y}_i)^2 = 775353{,}96$ при числе степеней свободы, равном $d.f. = n - k - 1$, $d.f. = 12$; *общая* — $\Sigma(y_i - \bar{y})^2 = 738126{,}94$, при числе степеней свободы $d.f. = n - 1$, $d.f. = 15$. Затем приводится средний квадрат отклонений: $s_1^2 = D_{\text{объясн}} : d.f._{\text{объясн}} = 662772{,}98 : 3 = 220924{,}3$; $s_2^2 = D_{\text{ост}} : d.f._{\text{ост}} = 75353{,}96 : 12 = 6279{,}5$. Далее указано их отношение, т.е. $s_1^2 / s_2^2 = F$-критерий. Наконец, указывается вероятность ошибочного решения, т.е. нулевого R^2, равная 0,000003171.

Три фактора, включенные в уравнение регрессии, объясняют 89,8% вариации уровня валового дохода, если рассматривать 16 хозяйств как генеральную совокупность, не считаясь с ее ограниченной численностью (некорректированный коэффициент детерминации равен 0,8979). Если же учесть конечность объема совокупности n, число факторов k, а также свойство метода, по которому по мере приближения числа k к числу n коэффициент детерминации автоматически приближается к 1 и достигает ее при $k = n - 1$ независимо от реальной роли факторов, то необходимо корректировать коэф-

фициент множественной детерминации на потерю степеней свободы вариации:

$$R^2_{\text{корр}} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right). \qquad (9.35)$$

Корректированный коэффициент детерминации всегда ниже, чем некорректированный, причем разность их значений тем меньше, чем меньше факторов входит в уравнение регрессии. Если из числа факторов исключить факторы, слабо связанные с результативным признаком (т.е. с низким значением β_j, например, $\beta_j < 0{,}1$), то некорректированный коэффициент детерминации немного уменьшится (он всегда уменьшается при исключении части факторов), но корректированный коэффициент может даже возрасти за счет уменьшения разности между R^2 и корректированным R^2. Что касается множественного коэффициента корреляции R^2, то программа «Microstat» рассчитывает его, как корень квадратный из некорректированного R^2, а другие программы, например «Statgraphics», — как корень квадратный из $R^2_{\text{корр}}$.

Для случая двух факторов коэффициент множественной детерминации легко вычисляется по рекуррентной формуле из парных коэффициентов детерминации:

$$R^2_{yx_1x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}}. \qquad (9.36)$$

Используя матрицу парных коэффициентов корреляции (табл. 9.10), получим:

$$R^2_{yx_1x_2} = \frac{0{,}687^2 + (-0{,}355)^2 - 2 \cdot 0{,}687 \cdot (-0{,}355) \cdot (-0{,}044)}{1 - (-0{,}044)^2} = 0{,}5765.$$

Таким образом, за счет вариации факторов x_1 и x_2 объясняется 57,65% общей вариации валового дохода с 1 га сельскохозяйственных угодий.

Вернемся к табл. 9.11. Дисперсионный анализ системы связей предназначен для оценки того, насколько надежно доказывают исходные данные наличие связи результативного признака со всеми факторами, входящими в уравнение. Для этого сравниваются дисперсии y — объясненная и остаточная: суммы соответствующих квадратов отклонений, приходящиеся на одну степень свободы вариации. Отношение дисперсии за счет факторов к остаточной дисперсии есть критерий Фишера F; в нашем примере он равен 35,18. Табличное, т.е. критическое, значение F для 3 и 12 степеней свободы при вероятности ошибочного отклонения нулевой гипотезы 0,01 составляет 5,95. Следовательно, вероятность ошибочного отклонения нулевой гипотезы намного меньше 0,01. Программа «Microstat» дает значение вероятности ошибочного отклонения нулевой гипотезы, т.е. вероятность случайного отклонения от нуля коэффициента детерминации при отсутствии связи в генеральной совокупности; она равна $3{,}17 \cdot 10^{-6}$, т.е. трем миллионным! Ясно, что эту ничтожную вероятность можно игнорировать и сделать вывод, что имеющаяся информация надежно свидетельствует о наличии связи.

Кроме показателя общей тесноты связи вариации результативного признака со всеми факторами, входящими в регрессионное уравнение, необходимы и показатели, измеряющие тесноту связи с каждым фактором. К таким показателям относятся коэффициенты раздельной детерминации.

Коэффициентом раздельной детерминации, обозначаемым далее d_j^2, называется произведение парного коэффициента корреляции фактора x_j на его β-коэффициент.

$$d_j^2 = r_{x_jy}\beta_j; \qquad \sum_{j=1}^{k} d_j^2 = R^2. \qquad (9.37)$$

Формула (9.36) дает еще один метод вычисления коэффициента множественной детерминации, используемый в некоторых программах для ЭВМ. В нашем примере получаем следующие значения коэффициентов раздельной детерминации:

$$d_1^2 = 0{,}687 \cdot 0{,}352 = 0{,}2418;$$

$$d_2^2 = -0{,}355 \cdot (-0{,}206) = 0{,}0731;$$

$$d_3^2 = 0{,}878 \cdot 0{,}664 = 0{,}5830.$$

Таким образом, за счет вариации x_1 объясняется 24,2% вариации y, за счет вариации x_2 — всего 7,3%; за счет вариации x_3 — более половины — 58,3% вариации уровня дохода. Сумма коэффициентов раздельной детерминации равна некорректированному коэффициенту R^2.

Недостатком коэффициентов раздельной детерминации является их гетерогенный характер: то, что они объединяют коэффициент *парной* корреляции, измеряющий нечистое влияние фактора, с β-коэффициентом, измеряющим условно-чистое влияние фактора, абстрагированное от влияния других факторов, входящих в уравнение связи. Из-за этого могут возникнуть неинтерпретируемые отрицательные величины коэффициентов d_j^2, если знаки парного коэффициента корреляции и β-коэффициента не совпадают при существенной взаимосвязи между факторами. Кроме того, сама идея о том, что совокупное влияние всех факторов равно сумме влияния каждого из них, противоречит системному подходу к исследованию корреляционных (стохастических) связей.

Рассмотрим разложение R^2 с учетом *системного эффекта*. Система факторов — это не простая их сумма, так как система предполагает внутренние связи, взаимодействие составляющих ее элементов. Действие системы не равно сумме воздействий составляющих ее элементов. К последним добавляется «системный эффект» («Emergency»). Методом, полностью отвечающим системному подходу, является метод разложения коэффициента множественной детерминации на сумму чистых влияний каждого фактора, выражаемую величинами β_1^2, и показатель влияния системного эффекта факторов η_s.

Поскольку расчетные значения результативного признака $\tilde{y}_i$ можно представить как $a + \sum_{j=1}^{k} b_j x_{ji}$, то вариацию $\tilde{y}_{ji}$ только за счет влияния фактора x_m можно представить при условии, что все остальные факторы, входящие в уравнение, закреплены на своих средних уровнях:

$$\tilde{y}(x_m)_i = a + b_m x_{m_i} + \sum_{j=1\ldots m-1,\, m+1,\ldots,k} b_j \bar{x}_j \,. \qquad (9.38)$$

Подставим в (9.38) значение фактора $x_{m_i} = \bar{x}_m + \Delta x_{m_i}$:

$$\tilde{y}(x_m)_i = a + b_m(\bar{x}_m + \Delta x_{m_i}) + \sum_{j=1\ldots m-1,\, m+1,\ldots,k} b_j \bar{x}_j =$$

$$= a + \sum_{j=1}^{k} b_j \bar{x}_j + b_m \Delta x_{m_i} = \bar{y} + b_m \Delta x_{m_i}.$$

Теперь измерим сумму квадратов отклонений y только за счет вариации признака x_m.

$$\sum_{i=1}^{n} (\tilde{y}(x_m)_i - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y} + b_m \Delta x_{m_i} - \bar{y})^2 = \sum_{i=1}^{n} (b_m \Delta x_{m_i})^2 =$$

$$= b_m^2 \sum_{i=1}^{n} (x_{m_i} - \bar{x}_m)^2 = b_m^2 n \sigma_{x_m}^2 . \qquad (9.39)$$

Мерой вариации результативного признака за счет изолированного влияния вариации фактора x_m является доля объясняемой этим влиянием вариации y. Соответственно получаем:

$$\frac{b_m^2 n \sigma_{x_m}^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{b_m^2 n \sigma_{x_m}^2}{n \sigma_y^2} = \left(b_m \frac{\sigma_{x_m}}{\sigma_y} \right)^2 = \beta_m^2 .$$

Сумма изолированных долей влияния каждого фактора в отдельности на вариацию y есть $\sum_{j=1}^{k} \beta_j^2$, а системный эффект

$$\eta_s = R^2 - \sum_{j=1}^{k} \beta_j^2 . \qquad (9.40)$$

Проведем разложение коэффициента множественной детерминации за счет вариации объясняющих переменных по данным нашего примера:

$x_1 : \beta_1^2 = 0{,}352^2 = 0{,}1239$, или 12,39%;

$x_2 : \beta_2^2 = (-0{,}206)^2 = 0{,}0424$, или 4,24%;

$x_3 : \beta_3^2 = 0{,}664^2 = 0{,}4409$, или 44,09%.

Суммарное влияние трех факторов составило: $\sum_{j=1}^{3} \beta_j^2 =$ = 60,72%. Системный эффект:

$$\eta_s = R^2 - \sum_{j=1}^{3} \beta_j^2 = 0{,}8979 - 0{,}6072 = 0{,}2907, \text{ или } 29{,}07\%.$$

Как видим, роль системного эффекта связей между факторами довольно велика: он на втором месте после влияния третьего фактора.

Системный эффект может, в свою очередь, быть разложен на влияние ковариации каждой пары факторов или на влияние совместной вариации отдельных групп факторов, если число последних велико. Если исследователь все же желает отказаться от выделения системного эффекта, свести коэффициент множественной детерминации к сумме эффектов отдельных факторов, можно распределить величину η_s пропорционально величинам β_j^2.

Программы анализа связей на ЭВМ обычно предусматривают вычисление *коэффициентов частной детерминации*. Они приведены в последней графе табл. 9.7. Коэффициент частной детерминации фактора x_m — это доля вариации y, дополнительно объясняемой при включении фактора x_m после остальных факторов в уравнение регрессии, в величине вариации y, не объясненной ранее включенными факторами. Наиболее ясно суть частных коэффициентов детерминации выражается формулой их расчета через коэффициенты множественной детерминации. *Частный коэффициент детерминации* для фактора x_m обозначим как

$$r^2_{yx_m(x_1 \ldots x_{m-1}, x_{m+1} \ldots x_k)}.$$

Тогда:

$$r^2_{yx_m(x_1 \ldots x_{m-1}, x_{m+1} \ldots x_k)} = \frac{R^2_y - R^2_{yx_1 \ldots x_{m-1}, x_{m+1} \ldots x_k}}{1 - R^2_{yx_1 \ldots x_{m-1}, x_{m+1} \ldots x_k}}. \qquad (9.41)$$

Здесь R^2_y — коэффициент детерминации для уравнения со всеми k факторами. Числитель (9.40) и есть дополнительно объясняемая часть вариации y при включении фактора x_m в уравнение после всех остальных факторов. В нашем примере, используя ранее рассчитанную величину $R^2_{yx_1x_2} = 0{,}5765$, а $R^2_{yx_1x_2x_3} = 0{,}8979$. Таким образом, при включении в анализ фактора x_3 получаем:

$$r^2_{yx_3(x_1x_2)} = \frac{R^2_y - R^2_{yx_1x_2}}{1 - R^2_{yx_1x_2}} = \frac{0{,}8979 - 0{,}5765}{1 - 0{,}5765} = \frac{0{,}3214}{0{,}4235} = 0{,}7589.$$

Некоторое расхождение в четвертой значащей цифре с табл. 9.7 объясняется округлением значений промежуточных расчетных показателей.

Следует усвоить, что коэффициенты частной детерминации — это доли от разных величин, поэтому они несравнимы; по этим долям нельзя судить о роли факторов. Их главное практическое значение — определить, имеет ли смысл добавить в уравнение регрессии новый фактор или нет. Если при его включении ранее необъясненная вариация уменьшится на 3/4, как в примере при введении фактора x_3, его включение оправдано; если же коэффициент частной детерминации мал, то дополнительный фактор включать не следует. Сумма частных коэффициентов детерминации смысла не имеет и растет с ростом числа факторов и ростом R^2 без ограничения.

При последовательном вводе факторов в уравнение регрессии объясняемая часть вариации результативного признака возрастает с каждым новым фактором, вводимым в уравнение. При вводе последнего фактора эта часть достигает величины R^2. Доли вариации y, объясняемые вводом каждого следующего фактора, и называют *коэффициентами последовательной детерминации*. Обозначим их как p_j^2. Для первого фактора этот коэффициент равен коэффициенту парной детерминации первого фактора, для второго — разности между коэффициентом детерминации при двух факторах и парным коэффициентом детерминации первого фактора и так далее. По данным нашего примера имеем:

$$p_1^2 = r^2_{yx_1} = 0{,}687^2 = 0{,}4720;$$

$$p_2^2 = R^2_{yx_1x_2} - r^2_{yx_1} = 0{,}5765 - 0{,}4720 = 0{,}1045;$$

$$p_3^2 = R^2_{yx_1x_2x_3} - R^2_{yx_1x_2} = 0{,}8979 - 0{,}5765 = 0{,}3214;$$

$$\sum_{j=1}^{k} p_j^2 = R^2.$$

Однако существенным недостатком такого способа разложения R^2 является зависимость величин p_j^2 от принятого порядка включения факторов в уравнение регрессии. Первый включаемый фактор «забирает в свою пользу» львиную долю системного эффекта, а последнему фактору остается ничтожная часть. Например, если переставить местами факторы x_1 и x_3, а также вычислить по рекуррентной формуле двухфакторный коэффициент детерминации $R^2_{yx_3x_2} = 0{,}8035$, то получим результаты, отличные от предыдущих:

$$p_1^2 \text{ (для фактора } x_3) = r^2_{yx_3} = 0{,}8782 = 0{,}7709;$$

$$p_2^2 \text{ (для фактора } x_2) = R^2_{yx_3x_2} - r^2_{yx_3} = 0{,}8035 - 0{,}7709 = 0{,}0326;$$

$$p_3^2 \text{ (для фактора } x_1) = R^2_{yx_1x_2x_3} - R^2_{yx_3x_2} = 0{,}8979 - 0{,}8035 = 0{,}0944.$$

Доля фактора x_3 возросла более чем вдвое, а доля фактора x_1 уменьшилась более чем втрое.

9.12. Вероятностные оценки параметров множественной регрессии и корреляции

Если показатели многофакторной системы связи используются как оценки генеральных параметров, экстраполируются на другие значения факторов, как это делается при прогнозировании, то значения параметров необходимо сопроводить вероятностными оценками, указать среднюю ошибку и доверительные границы параметра с заданной вероятностью. Для парной корреляции эта проблема изложена в подразд. 9.5. Там же приводятся формулы средних ошибок репрезентативности для специфических параметров многофакторной системы.

Средняя ошибка условно-чистого коэффициента регрессии b_p для фактора x_p, обозначаемая m_{b_p}, имеет вид:

$$m_{b_p} = \frac{s_{y_{ост}}}{s_{x_p}\sqrt{n}} \sqrt{\frac{1}{1 - R^2_{x_p, x_1 \ldots x_{p-1}, x_{p+1} \ldots x_k}}}, \qquad (9.42)$$

где $s_{y_{ост}}$ — оценка остаточного (не объясненного факторами) среднего квадратического отклонения результативного признака с учетом степеней свободы вариации:

$$s_{y_{ост}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{n-k-1}},$$

s_{x_p} — оценка среднего квадратического отклонения признака x_p:

$$s_{x_p} = \sqrt{\frac{\sum_{i=1}^{n} (x_{p_i} - \bar{x}_p)^2}{n-1}};$$

$R^2_{x_p, x_1 \ldots x_{p-1}, x_{p+1} \ldots x_k}$ — коэффициент множественной детерминации для фактора x_p; $1 - R^2_{x_p, x_1 \ldots x_{p-1}, x_{p+1} \ldots x_k}$ — доля вариации фактора x_p, связанная с вариацией других факторов.

Например, для фактора x_1 имеем:

$$s_{y_{ост}} = 79{,}24;$$

$$s_{x_1} = 34{,}6.$$

Коэффициент $R^2_{x_1x_2x_3} = 0{,}2433$ вычислен по рекуррентной формуле по данным табл. 9.10. Отсюда:

$$m_{b_1} = \frac{79{,}24}{34{,}6\sqrt{16}} \sqrt{\frac{1}{1 - 0{,}2433}} = 0{,}6582.$$

Отношение величины коэффициента регрессии к его средней ошибке есть t-критерий Стьюдента. В данном случае имеем: $t = b_1/m_{b_1} = 2{,}26/0{,}6582 = 3{,}43$. Критическое значение t для вероятности нулевой гипотезы 0,01 при 12 степенях свободы равно 3,05. Следовательно, надежно установлено, что генеральное значение коэффициента b_1 не является нулевым, влияние (условно-чистое) фактора x_1 на вариацию y существенно.

Доверительные границы коэффициента регрессии b_1 с вероятностью 0,95, для которой значение критерия Стьюдента равно 2,18,составляют $2{,}26 \pm 2{,}18 \cdot 0{,}658$, или от 0,826 до 3,694.

Столь широкие границы объясняются малой численностью единиц совокупности. Из (9.42) следует, что при росте объема совокупности в q раз ошибка коэффициента регрессии, как и ошибка выборочной оценки средней величины, уменьшится в $\sqrt{q}$ раз. При 400 единицах совокупности ошибка была бы меньше в 5 раз.

Если значение t-критерия оказывается ниже критического для вероятности нулевой гипотезы 0,05, влияние фактора считается недоказанным надежно, и при работе программ ЭВМ с отсевом несущественных факторов по t-критерию данный фактор автоматически исключается из уравнения регрессии.

Средняя ошибка оценки коэффициента множественной корреляции m_R определяется по формуле

$$m_R = \sqrt{\frac{1-R^2}{n-k-1}}. \qquad (9.43)$$

Оценка существенности и расчет доверительных границ генерального коэффициента корреляции осуществляются так же, как и для коэффициента регрессии. Если значение R близко к 1, необходимо использовать преобразование Фишера, рассмотренное в подразд. 9.5. Имеются также специальные таблицы критических значений коэффициента корреляции для заданного числа степеней свободы и вероятности нулевой гипотезы (табл. П.5 приложения).

9.13. Корреляционно-регрессионные модели и их применение в анализе и прогнозе

Корреляционно-регрессионной моделью (*КРМ*) системы взаимосвязанных признаков является такое уравнение регрессии, которое включает основные факторы, влияющие на вариацию результативного признака, обладает высоким (не ниже 0,5) коэффициентом детерминации и коэффициентами регрессии, интерпретируемыми в соответствии с теоретическим знанием о природе связей в изучаемой системе.

Приведенное определение КРМ включает достаточно строгие условия: далеко не всякое уравнение регрессии можно считать моделью. В частности, полученное выше по 16 хозяйствам уравнение не отвечает последнему требованию из-за противоречащего экономике сельского хозяйства знака при факторе x_2 — доля пашни. Однако в учебных целях будем рассматривать его как модель.

Теория и практика выработали ряд рекомендаций для построения корреляционно-регрессионной модели.

1. Признаки-факторы должны находиться в причинной связи с результативным признаком (следствием). Поэтому недопустимо, например, в модель себестоимости y вводить в качестве одного из факторов x_j коэффициент рентабельности, хотя включение такого «фактора» значительно повысит коэффициент детерминации.

2. Признаки-факторы не должны быть составными частями результативного признака или его функциями.

3. Признаки-факторы не должны дублировать друг друга, т.е. быть коллинеарными (с коэффициентом корреляции более 0,8). Так, не следует в модель производительности труда включать энерго- и фондовооруженность рабочих, поскольку эти факторы тесно связаны друг с другом в большинстве объектов.

4. Не следует включать в модель факторы разных уровней иерархии, т.е. фактор ближайшего порядка и его субфакторы. Например, в модель себестоимости зерна не следует включать и урожайность зерновых культур, и дозу удобрений под них или затраты на обработку гектара, показатели качества семян, плодородия почвы, т.е. субфакторы самой урожайности.

5. Желательно, чтобы для результативного признака и факторов соблюдалось единство единицы совокупности, к которой они отнесены. Например, если y — валовой доход предприятия, факторы должны относиться к предприятию: стоимость производственных фондов, уровень специализации, численность работников и т.д. Если же y — средняя зарплата рабочего на предприятии, то факторы должны относиться к рабочему: разряд или классность, стаж работы, возраст, уровень образования, энерговооруженность и т.д. Правило это некатегорическое, в модель заработной платы рабочего можно включить, к примеру, и уровень специализации предприятия. Вместе с тем нельзя забывать о предыдущей рекомендации.

6. Математическая форма уравнения регрессии должна соответствовать логике связи факторов с результатом в реальном объекте. Например, такие факторы урожайности, как дозы разных удобрений, уровень плодородия, число прополок и т.п., создают прибавки величины урожайности, малозависящие друг от друга; урожайность может существовать и без любого из этих факторов. Такому характеру связей отвечает аддитивное уравнение регрессии:

$$\hat{y} = a + b_1 x_2 + b_2 x_2 + \ldots + b_{k_1} x_{k_1}.$$

Наоборот, если y — объем валовой продукции завода, x_1 — число работников, x_2 — стоимость основных производственных фондов, x_3 — затраты на энергию, топливо, сырье и материалы (комплектующие изделия), то результат без любого из факторов не существует, поэтому большинство экономистов-статистиков строят КРМ, называемую *производственной функцией* (что весьма неудачно терминологически) в мультипликативной форме:

$$y = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot x_3^{b_3} \ldots x_k^{b_k}, \qquad (9.44)$$

где коэффициенты b_j соответствуют коэффициентам эластичности факторов при стремлении прироста фактора к бесконечно малой величине: $b_j \to e_j$ при $\Delta x_j \to 0$.

Для конечных приростов факторов коэффициенты уравнения (9.43) не равны коэффициентам эластичности, как иногда утверждается в литературе.

Уравнение (9.44) линеаризуется логарифмированием:

$$\ln \hat{y} = b_1 \ln x_1 + b_2 \ln x_2 + b_3 \ln x_3 + \ln a$$

и далее оценки параметров находятся как для линейной модели.

7. Принцип простоты: предпочтительнее модель с меньшим числом факторов при том же коэффициенте детерминации или даже при несущественно меньшем.

Для анализа степени эффективности управления производством можно использовать сравнение единиц совокупности по показателям отклонений результативного признака от средней величины и от значения, рассчитанного по уравнению регрессии

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \qquad (9.45)$$

Первое слагаемое в правой части равенства — это отклонение, которое возникает за счет отличия индивидуальных значений факторов y данной единицы совокупности от их средних значений по совокупности. Его можно назвать *эффектом факторообеспеченности*. Второе слагаемое — отклонение, которое возникает за счет не входящих в модель факторов и отличия индивидуальной эффективности факторов у данной единицы совокупности от средней эффективности факторов в совокупности, измеряемой коэффициентами условно-чистой регрессии. Его можно назвать *эффектом фактороотдачи*.

Таблица 9.12

Анализ факторообеспеченности и фактороотдачи по регрессионной модели уровня валового дохода

Номер хозяйства	Уровни дохода, руб./га		Отклонения уровней, руб./га		Анализ связи отклонений			
	y_i	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$y_i - \hat{y}_i$	ранг по $\hat{y}_i - \bar{y}$	ранг по $y_i - \hat{y}_i$	разность Δp	Δ_p^2
1	704	733	+128	−29	13	7	6	36
2	293	370	−235	−77	2	3	−1	1
3	346	433	−172	−87	3	2	1	1
4	420	477	−128	−57	5	4	1	1
5	691	566	−39	+125	8	16	−8	64
6	679	591	−14	+88	10	14	−4	16
7	457	474	−131	−17	4	8	−4	16
8	503	548	−57	−45	7	5	2	4
9	314	272	−333	+42	1	12	−11	121
10	803	833	+228	−30	14	6	8	64
11	691	631	+26	+60	11	13	−2	4
12	775	674	+69	+101	12	15	−3	9
13	584	573	−32	+11	9	9	0	0
14	504	493	−112	+11	6	10	−4	16
15	777	899	+294	−122	15	1	14	196
16	1138	1110	+505	+28	16	11	5	25
Итого	605	605	0	0	—	—	0	574

Пример. Рассмотрим расчет и анализ отклонений по ранее построенной модели уровня валового дохода в 16 хозяйствах. Знаки тех и других отклонений 8 раз совпадают и 8 раз не совпадают. Коэффициент корреляции рангов отклонений двух видов составил 0,156. Это означает, что связь вариации факторообеспеченности с вариацией фактороотдачи слабая, несущественная (табл. 9.12).

Обратим внимание на хозяйство № 15 с высокой факторообеспеченностью (15-е место) и самой худшей фактороотдачей (1-й ранг), из-за которой хозяйство недополучило по 122 руб. дохода с 1 га. Напротив, хозяйство № 5 имеет фак-

торообеспеченность ниже средней, но благодаря более эффективному использованию факторов получило на 125 руб. дохода с 1 га больше, чем было бы получено при средней по совокупности эффективности факторов. Более высокая эффективность фактора x_1 (затраты труда) может означать более высокую квалификацию работников и бо́льшую заинтересованность в качестве выполняемой работы. Более высокая эффективность фактора x_3 с точки зрения доходности может заключаться в высоком качестве молока (жирность, охлажденность), благодаря которому оно реализовано по более высоким ценам. Коэффициент регрессии при x_2, как уже отмечено, экономически не обоснован.

Использование регрессионной модели для прогнозирования состоит в подстановке в уравнение регрессии ожидаемых значений факторных признаков для расчета точечного прогноза результативного признака или (и) его доверительного интервала с заданной вероятностью, как уже сказано в 9.6. Сформулированные там же ограничения прогнозирования по уравнению регрессии сохраняют свое значение и для многофакторных моделей. Кроме того, необходимо соблюдать системность между подставляемыми в модель значениями факторных признаков.

Формулы расчета средних ошибок оценки положения гиперплоскости регрессии в заданной многомерной точке и для индивидуальной величины результативного признака весьма сложны, требуют применения матричной алгебры и здесь не рассматриваются. Средняя ошибка оценки значения результативного признака, рассчитанная по программе ПЭВМ «Microstat» и приведенная в табл. 9.7, равна 79,2 руб. на 1 га. Это лишь среднее квадратическое отклонение фактических значений дохода от расчетных по уравнению, не учитывающее ошибки положения самой гиперплоскости регрессии при экстраполяции значений факторных признаков. Поэтому ограничимся точечными прогнозами в нескольких вариантах (табл. 9.13).

Для сравнения прогнозов с базисным уровнем средних по совокупности значений признаков введена первая строка таблицы. Краткосрочный прогноз рассчитан на малые изменения факторов за короткое время и снижение трудообеспеченности.

Таблица 9.13

Прогнозы валового дохода по регрессионной модели

Вариант прогноза	Возможные значения факторов			Ожидаемое по модели значение результативного признака	
	x_1	x_2	x_3	$\tilde{y}$, руб./га	в % к базисному
1. Фактический (базисный)	218,2	54,1	3520	604,9	100
2. Краткосрочный	200	50	3600	594,1	98,2
3. Долгосрочный А	220	55	4000	684,2	113,1
4. « Б	250	55	4500	834,9	138,0
5. «Идеальный жених»	276	35,1	5526	1149,8	190,1

Результат неблагоприятен: доход снижается. Долгосрочный прогноз А — «осторожный», он предполагает весьма умеренный прогресс факторов и соответственно небольшое увеличение дохода. Вариант Б — «оптимистический», рассчитан на существенное изменение факторов. Вариант 5 построен по способу, которым Агафья Тихоновна в комедии Н. В. Гоголя «Женитьба» мысленно конструирует портрет «идеального жениха»: нос взять от одного претендента, подбородок от другого, рост от третьего, характер от четвертого, вот если бы соединить все нравящиеся ей качества в одном человеке, она бы не колеблясь вышла замуж. Так и при прогнозировании мы объединяем лучшие (с точки зрения модели дохода) наблюдаемые значения факторов: берем значение x_1 от хозяйства № 10, значение x_2 от хозяйства № 2, значение x_3 от хозяйства № 16. Все эти значения факторов уже существуют реально в изучаемой совокупности, они не «ожидаемые», не «взятые с потолка». Это хорошо. Однако могут ли эти значения факторов сочетаться в одном предприятии, системны ли эти значения? Решение данного вопроса выходит за рамки статистики, оно требует конкретных знаний об объекте прогнозирования.

Если, кроме количественных факторов, при многофакторном регрессионном анализе в уравнение включается и неколичественный, то применяют следующую методику: наличие неколичественного фактора у единиц совокупности обозначают единицей, его отсутствие — нулем, т.е. вводят так назы-

ваемую *фиктивную переменную* $u = \begin{cases} 1 \\ 0 \end{cases}$. Если таких переменных, или градаций неколичественного фактора, несколько, в уравнение регрессии вводится несколько фиктивных переменных. Пусть имеются три количественных фактора урожайности (x_1, x_2, x_3) и три природных зоны. В ПЭВМ вводятся переменные в порядке их принадлежности к той или иной зоне (табл. 9.14).

Линейное уравнение регрессии будет иметь вид:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4u_1 + b_5u_2. \qquad (9.46)$$

Величина коэффициента b_4 означает, что все единицы II зоны при тех же значениях количественных факторов, как и

Таблица 9.14

Зоны	Результативный признак y_i	Количественные факторы			Фиктивные переменные	
		x_1	x_2	x_3	u_1	u_2
I	y_1	x_{11}	x_{21}	x_{31}	0	0
	y_2	x_{12}	x_{22}	x_{32}	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	y_{n_1}	x_{n_1}	x_{2n_1}	x_{3n_1}	0	0
II	y_{n_1+1}	x_{1n_1+1}	x_{2n_1+1}	x_{3n_1+1}	1	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	$y_{n_1+n_2}$	$x_{1n_1+n_2}$	$x_{2n_1+n_2}$	$x_{3n_1+n_2}$	1	0
III	$y_{n_1+n_2+1}$	$x_{1n_1+n_2+1}$	$x_{2n_1+n_2+1}$	$x_{3n_1+n_2+1}$	0	1
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	$y_{n_1+n_2+n_3}$	$x_{1n_1+n_2+n_3}$	$x_{2n_1+n_2+n_3}$	$x_{3n_1+n_2+n_3}$	0	1

единицы I зоны, будут в среднем иметь значение $\hat{y}$ на b_4 больше (или меньше, если $b_4 < 0$), чем единицы совокупности I зоны. Величина b_5 означает то же для единиц совокупности III зоны. Иначе говоря, мы получаем сразу три зональных регрессионных модели:

$$\text{I: } \hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3; \ (u_1 = 0; \ u_2 = 0);$$
$$\text{II: } \hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4u_1; \ (u_2 = 0);$$
$$\text{III: } \hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + b_5u_2; \ (u_1 = 0).$$

Число фиктивных переменных должно быть на единицу меньше числа градаций качественного (неколичественного) фактора. С помощью данного приема можно измерять влияние уровня образования, местожительства, типа жилища и других социальных или природных, неизмеряемых количественно факторов, изолируя их от влияния количественных факторов.

РЕЗЮМЕ

Связи, которые проявляются не в каждом отдельном случае, а лишь в совокупности данных, называются статистическими. Они выражаются в том, что при изменении значения фактора x изменяется и условное распределение результативного признака y: разным значениям одной переменной (фактора x) соответствуют разные распределения другой переменной (результата y).

Корреляционная связь — частный случай статистической связи, при котором разным значениям одной переменной x соответствуют разные средние значения переменной y.

Корреляционная связь предполагает, что изучаемые переменные имеют количественное выражение.

Статистическая связь — более широкое понятие, оно не включает ограничений на уровень измерения переменных. Переменные, связь между которыми изучается, могут быть как количественными, так и неколичественными.

Статистические связи отражают сопряженность в изменении признаков x и y, которая может быть вызвана не причинными отношениями, а так называемой ложной корреляцией. Например, в совместных изменениях x и y обнаруживается определенная закономерность, но она вызвана не влиянием

x на y, а тем, что обе переменные x и y изменяются под влиянием общей причины z. Следует проявлять осторожность при интерпретации результатов измерения статистических связей. Когда говорится, что изменение y на 70% зависит от изменения фактора x, нужно понимать условность такого вывода и ставить под сомнение как сам вывод о зависимости, так и цифру — в данном случае 70%, которая отражает не только влияние изучаемого фактора, но и всего комплекса факторов, связанных с ним.

Показатели корреляции измеряют тесноту связи между признаками. Все показатели корреляции изменяются по абсолютной величине в интервале [0; 1].

Коэффициент парной корреляции — мера тесноты линейной связи между двумя переменными x и y. Линейная связь может быть либо прямой, $r_{yx} > 0$, либо обратной, $r_{yx} < 0$, так что $-1 \leqslant r_{yx} \leqslant 1$. Коэффициент корреляции — это симметричная мера связи, т.е. $r_{yx} = r_{xy}$. Квадрат коэффициента корреляции называется парным коэффициентом детерминации.

Коэффициент частной корреляции измеряет чистую (частную) корреляцию между двумя переменными при погашении связи с другими переменными. Коэффициент частной корреляции также является мерой линейной связи и принимает значения от -1 до 1.

Коэффициент частной детерминации переменной x_k — это доля дисперсии y, дополнительно объясненной включением переменной x_k, в величине дисперсии y, не объясненной переменными, ранее включенными в анализ $(x_1, \ldots, x_{k-1})$.

Коэффициент множественной корреляции измеряет связь между y и всеми учтенными признаками-факторами: $-1 \leqslant \leqslant R^2_{x,y,z} \leqslant 1$.

Коэффициент множественной детерминации показывает, какую часть дисперсии y объясняют учтенные в анализе признаки-факторы: $0 \leqslant R^2_{x,y,z} \leqslant 1$.

Математическое описание зависимости изменений переменной y в среднем от изменений переменной x называется уравнением парной регрессии. Чаще всего используется линейное уравнение парной регрессии: $\hat{y} = a + bx$. Знак при коэффициенте регрессии b соответствует направлению зависимости y от x: $b > 0$ — зависимость прямая, $b < 0$ — зависимость обратная. Коэффициент регрессии b_{yx} измеряет силу зависимости y от x. Это асимметричный показатель, т.е. $b_{yx} \neq b_{xy}$.

Интерпретация свободного члена a зависит от того, имеется ли в исходных данных нулевое значение x (и возможно ли оно).

Математическое описание корреляционной зависимости результативной переменной от нескольких факторных переменных называется уравнением множественной регрессии. Параметры уравнения регрессии оцениваются методом наименьших квадратов (МНК). Уравнение регрессии должно быть линейным по параметрам.

Если уравнение регрессии отражает нелинейность связи между переменными, то регрессия приводит к линейному виду (линеаризуется) путем замены переменных или их логарифмирования.

Вводя в уравнение регрессии фиктивные переменные, можно учесть влияние неколичественных переменных, изолируя их от влияния количественных факторов.

Если коэффициент детерминации близок к единице, то с помощью уравнения регрессии можно предсказать, каким будет значение зависимой переменной для того или иного ожидаемого значения одной или нескольких независимых переменных.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Елисеева И. И.* Статистические методы измерения связей. — Л.: Изд-во Ленингр. ун-та, 1982.
2. *Елисеева И. И., Рукавишников В. О.* Логика прикладного статистического анализа. — М.: Финансы и статистика, 1982.
3. *Крастинь О. П.* Разработка и интерпретация моделей корреляционных связей в экономике. — Рига: Зинатне, 1983.
4. *Кулаичев А. П.* Методы и средства анализа данных в среде Windows. Stadia 6.0. — М.: НПО «Информатика и компьютеры», 1996.
5. *Статистическое* моделирование и прогнозирование: Учеб. пособие / Под ред. А. Г. Гранберга. — М.: Финансы и статистика, 1990.
6. *Ферстер Э., Ренц Б.* Методы корреляционного и регрессионного анализа. Руководство для экономистов: Пер. с нем. — М.: Финансы и статистика, 1983.

Глава
10

СИСТЕМЫ РЕГРЕССИОННЫХ УРАВНЕНИЙ

10.1. Понятие о системах регрессионных уравнений

Выше были последовательно рассмотрены методы анализа связи одного результативного показателя с одним фактором (парная корреляция и парная регрессия), затем — связь одного результативного показателя с несколькими факторами (множественная корреляция и множественная регрессия). В реальных экономических, технологических, природных и социальных системах многие результативные и факторные признаки взаимосвязаны. В этом случае статистическими методами определяется не один результативный признак, а несколько, каждый из которых имеет ряд факторов, причем сами результативные признаки также связаны друг с другом.

Например, для успешной деятельности предприятия очень важно определить взаимосвязанные уровни производительности труда (y_1) и его оплаты (y_2). На каждый из них влияет ряд факторных признаков: так, на y_1 влияют энерго- и фондовооруженность работников, стаж работы, квалификация работников, а также уровень производительности труда. Одни факторы являются общими для y_1 и y_2, другие — различными. Подобную систему можно изобразить в виде графа связей (рис. 10.1).

Здесь x_1, x_2, ..., x_k — факторные признаки (независимые переменные), считающиеся известными; y_1, y_2 — результативные, моделируемые признаки. Стрелками показано влияние одних признаков на другие. Для каждой конкретной системы, т.е. конкретной задачи анализа, признаки, подлежащие определению («игреки»), принято называть *эндогенными*, или внут-

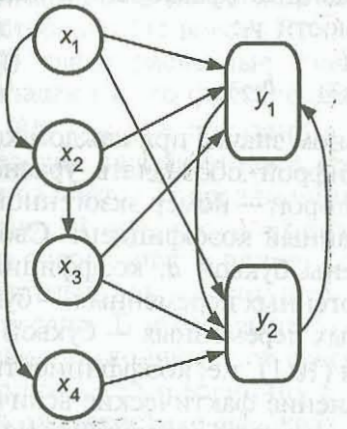


Рис. 10.1. Граф связей

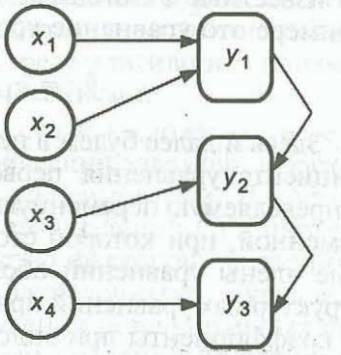


Рис. 10.2. Граф связей рекуррентной системы

ренними, а признаки, считающиеся для данной задачи известными (заданными), т.е. «иксы» — *экзогенными*, или внешними. Признак, эндогенный для одной задачи, может являться экзогенным в другой задаче и наоборот.

Важнейшей особенностью, определяющей методику исследования систем с несколькими результативными признаками, является характер связи между ними. Если эта связь односторонняя, т.е. один результативный признак в данной задаче играет только роль следствия, *такая система называется рекуррентной* (или *рекурсивной*). Граф связей рекуррентной системы изображен на рис. 10.2.

Примером такой системы может быть связь результативных признаков в производстве сельскохозяйственной культуры: y_1 — урожайность; y_2 — себестоимость центнера; y_3 — рентабельность отрасли в сельскохозяйственном предприятии или регионе. В рекуррентной системе отсутствует обратное влияние последующих по графу результативных признаков на предыдущие. Вследствие этого методика моделирования рекуррентной системы меньше отличается от обычной методики регрессионного анализа.

Эта методика такова: сначала обычным МНК решается уравнение того результативного признака, который занимает «верхнее» место в графе связей, т.е. на который влияют толь-

ко известные экзогенные переменные (факторы). В данном примере это уравнение урожайности y_1:

$$\hat{y}_1 = a_1 + b_{11}x_1 + b_{12}x_2. \qquad (10.1)$$

Здесь и далее будем в подписном значке при каждом коэффициенте уравнения первой цифрой обозначать уравнение (определяемую переменную), а второй — номер экзогенной переменной, при которой стоит данный коэффициент. Свободные члены уравнений обозначены буквой *a*, коэффициенты структурных уравнений при экзогенных переменных — буквой *b*, коэффициенты при эндогенных переменных — буквой *c*.

Получив решение уравнения (10.1), т.е. коэффициенты a_1, b_{11}, b_{12}, подставляем в это уравнение фактические величины экзогенных переменных для всех единиц совокупности и получаем расчетные значения эндогенной переменной $\hat{y}_{1i}$.

Затем решается второе уравнение рекуррентной системы — в примере это уравнение себестоимости, y_2. Для его решения используются значение экзогенного признака x_{3i} и расчетные значения $\hat{y}_{4i}$, полученные на предыдущем этапе.

Аналогично при решении третьего рекуррентного уравнения следует использовать факторные значения экзогенной переменной x_{4i} и *расчетные значения* предыдущей эндогенной $\hat{y}_{2i}$. В результате будет получено численное выражение коэффициентов третьего уравнения— рентабельности отрасли.

10.2. Проблемы решения систем взаимосвязанных уравнений

В чем заключается необходимость использовать при решении рекуррентных уравнений не фактические значения «вышележащих», т.е. предшествующих по графу связей, играющих роль причины эндогенных переменных, а их расчетные значения, полученные из решения предыдущего уравнения? Разобраться в этой проблеме тем более необходимо, что она относится не только к рекуррентным, но и ко всем иным системам взаимосвязанных регрессионных уравнений. Если бы в число экзогенных переменных, входящих в правые части уравнений, входили все факторы, определяющие вариацию каждой эндогенной переменной, т.е. имели бы место

функциональные связи, а не стохастические, проблемы не существовало бы вовсе. При функциональной (полной, жесткой) связи расчетные значения результативного признака совпадают с его фактическими значениями.

Но, как уже показано в гл. 9, связи в сложных системах массовых явлений имеют стохастический характер, в частности являются корреляционными зависимостями. На результативный признак, на вариацию его значений у разных единиц совокупности влияет множество факторов, частично не известных или не могущих быть включенными в уравнение регрессии. В итоге расчетные значения результативных признаков отклоняются от фактических значений на некоторую стохастическую величину ε_{y_i}, часто называемую в математической статистике *ошибкой*. Этот термин следует признать методически неудачным: у изучающих статистику слово «ошибка» часто ассоциируется с искажениями в учете, преднамеренными или случайными ошибками регистрации, т.е. с недостоверной информацией.

Проблема, которая здесь рассматривается, состоит в том, что при корреляционно-регрессионной связи последующей эндогенной переменной y_2 с предыдущей y_1, если использовать ее фактические значения, эндогенная переменная y_2 окажется коррелированной не только с известными и входящими в уравнение y_1 экзогенными переменными, но и прочими факторами вариации y_{1i}. Как говорят математики, переменная будет «коррелирована и с ошибками», с неизвестными значениями $\varepsilon_{y_{1i}}$. Эта корреляция последующих эндогенных переменных с неизвестными факторами вариации предыдущих эндогенных приводит к нарушению предпосылок или условий, при которых обычный МНК дает несмещенные и состоятельные оценки результативных признаков. С точки зрения материального содержания изучаемых признаков это означает, что при включении в регрессионное уравнение для последующей эндогенной переменной фактических значений предыдущей эндогенной переменной на вариацию последующей переменной окажут влияние такие факторы предыдущей, которые никакого отношения, по существу, к последующей эндогенной переменной не имеют. Пусть, например, y_1 — это урожайность сельскохозяйственной культуры, а ее известные эндогенные факторы x_{11}; x_{12}... — это доза различных удобре-

ний, число поливов, прополок, затраты труда на гектар посевов и т.п., y_2 — себестоимость центнера или тонны данной культуры; ее экзогенные факторы $x_{21}; x_{22}...$ — это цена на горючее, ставки оплаты часа труда, уровень накладных расходов на 1 га, а также y_1 — урожайность.

Однако на урожайность в каждом хозяйстве, т.е. на фактические значения y_{1j}, повлияют не только входящие в уравнение факторы, но и такой, как правило, неизвестный фактор, как сумма атмосферных осадков на полях данного j-го хозяйства. Если в расчет параметров регрессионного уравнения себестоимости $\hat{y}_2$ включить фактические величины урожайности по хозяйствам (плюс экзогенные факторы y_2), то через вариацию урожайностей y_{1j} на параметры уравнения себестоимости станет влиять и сумма осадков на полях хозяйств, точнее — вариация этой суммы осадков за сезон по разным хозяйствам. Это исказит оценки параметров уравнения себестоимости и ее расчетные значения; ведь очевидно, что естественные осадки — даровой фактор урожайности, который к себестоимости в отличие от доз удобрений никакого отношения иметь не должен.

Несмещенность оценки какого-либо параметра означает, что математическое ожидание его выборочных оценок (т.е. сумма произведений возможных выборочных оценок на их вероятности) равно значению параметра в генеральной совокупности. Метод наименьших квадратов обеспечивает несмещенность оценок параметров, если между входящими в правую часть уравнения переменными и не входящими в нее факторами («ошибками») корреляция отсутствует. По отношению к известным экзогенным переменным уравнения себестоимости это условие соблюдено: $x_{21}; x_{22}...$ — цена на горючее, ставки тарифной сетки и т.д., конечно, не коррелированы с суммой атмосферных осадков! Иное дело стоящая в правой же части уравнения себестоимости эндогенная переменная y_1 — урожайность, явно коррелированная с осадками, но лишь тогда, когда мы включаем ее фактические значения по хозяйствам. В таком решении оценки параметров уравнения себестоимости окажутся *смещенными*. Если же в правую часть уравнения регрессии себестоимости включить значения расчетной урожайности $\hat{y}_{1j}$, то влияние вариации осадков исключается, поскольку значения $\hat{y}_{1j}$ получены при решении

первого рекуррентного уравнения, в правую часть которого величина осадков не входила, а входящие в эту часть объясняющие переменные $x_{11}; x_{12}...$, в свою очередь, не коррелированы с суммой осадков. Поэтому при замене фактических значений предыдущей эндогенной переменной y_1 на ее расчетные значения $\hat{y}_1$ в уравнении последующей эндогенной переменной $\hat{y}_2$ МНК дает несмещенные и состоятельные оценки параметров.

Методика исключения влияния не входящих в уравнение факторов одной переменной на другую различается в зависимости от типа системы уравнений, но суть дела остается той же. Что касается систем уравнений рекуррентного типа, то, кроме рассмотренной особенности, последующий алгоритм решения ничем не отличается от методики, изложенной в гл. 9, для множественных корреляционно-регрессионных связей.

10.3. Преобразование структурных уравнений в приведенные и их идентификация

Рассмотрим более подробно методы исследования решения таких систем регрессионных уравнений, в которых связь между эндогенными переменными является двусторонней, т.е. каждая из них влияет на вариацию другой переменной и, в свою очередь, зависит от вариации другой (или нескольких других) эндогенной переменной (см. рис. 10.1).

Запишем уравнение этой системы, придерживаясь ранее принятой системы обозначения и нумерации переменных:

$$\hat{y}_1 = a_1 + b_{11}x_1 + b_{12}x_2 + b_{13}x_3 + c_{12}y_2,$$
$$\hat{y}_2 = a_2 + b_{21}x_1 + b_{23}x_3 + b_{24}x_4 + c_{21}y_1. \qquad (10.2)$$

Система уравнений (10.2), соответствующая структуре связей, изображенной в виде графа связи, называется *системой структурных уравнений*. Каждое из структурных уравнений рассматриваемой системы имеет по пять параметров, включая свободные члены. Некоторые факторы, т.е. экзогенные переменные, являются общими для двух уравнений. Таких факторов два — x_1 и x_3. А часть экзогенных переменных входит только в одно из уравнений — это x_2 и x_4. В каждом уравнении отсутствует по одному из экзогенных факторов и

присутствует в правой части по одной эндогенной переменной. Ввиду рассмотренного свойства систем регрессионных уравнений решать их непосредственно в структурной форме не следует из-за потери свойств несмещенности и состоятельности оценок параметров.

Решать можно лишь такое уравнение, которое в правой части не содержит эндогенных переменных. Для того чтобы привести систему к такой форме, следует выразить значение каждой из эндогенных переменных через другую эндогенную и экзогенные переменные и подставить это выражение в другое уравнение. Тогда в каждом уравнении останется по одной эндогенной переменной, которую следует «поместить» в левой части уравнения, оставив в правой части только экзогенные переменные. Такие уравнения называют *приведенными*. Подставив $\hat{y}_1$ в уравнение $\hat{y}_2$, получим:

$$\hat{y}_2 = a_2 + b_{21}x_1 + b_{23}x_3 + b_{24}x_4 + c_{21}(a_1 + b_{11}x_1 + b_{12}x_2 + b_{13}x_3 + c_{12}\hat{y}_2).$$

Затем, объединив однородные члены и перенеся $c_{12}\hat{y}_2$ в левую часть уравнения, получим:

$$(1 - c_{21}c_{12})\hat{y}_2 = (a_2 + c_{21}a_1) + (b_{21} + c_{21}b_{11})x_1 + c_{21}b_{12}x_2 + (b_{23} + c_{21}b_{13})x_3 + b_{24}x_4.$$

Разделив обе части уравнения на коэффициент при $\hat{y}_2$, получим:

$$\hat{y}_2 = \frac{a_2 + c_{21}a_1}{1 - c_{21}c_{12}} + \frac{b_{21} + c_{21}b_{11}}{1 - c_{21}c_{12}}x_1 + \frac{c_{21}b_{12}}{1 - c_{21}c_{12}}x_2 + \frac{b_{23} + c_{21}b_{13}}{1 - c_{21}c_{12}}x_3 + \frac{b_{24}}{1 - c_{21}c_{12}}x_4.$$

Для краткости записи обозначим коэффициенты регрессии приведенных уравнений буквой дельта (δ) с соответствующими подписными значками, а свободные члены — буквой альфа (α).

$$\hat{y}_2 = \alpha_2 + \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3 + \delta_{24}x_4. \qquad (10.3)$$

Аналогично в уравнение $\hat{y}_1$ подставим значение $\hat{y}_2$ и получим:

$$\hat{y}_1 = a_1 + b_{11}x_1 + b_{12}x_2 + b_{13}x_3 + c_{12}(a_2 + b_{21}x_1 + b_{23}x_3 + b_{24}x_4 + c_{21}\hat{y}_1).$$

Объединив однородные члены и перенеся $c_{21}\hat{y}_1$ в левую часть, получим:

$$(1 - c_{12}c_{21})\hat{y}_1 = (a_1 + c_{12}a_2) + (b_{11} + c_{12}b_{21})x_1 + b_{12}x_2 + (b_{13} + c_{12}b_{23})x_3 + c_{12}b_{24}x_4.$$

Разделив все члены на $(1 - c_{12}c_{21})$, имеем:

$$\hat{y}_1 = \frac{a_1 + c_{12}a_2}{1 - c_{12}c_{21}} + \frac{b_{11} + c_{12}b_{21}}{1 - c_{12}c_{21}}x_1 + \frac{b_{12}}{1 - c_{12}c_{21}}x_2 + \frac{b_{13} + c_{12}b_{23}}{1 - c_{12}c_{21}}x_3 + \frac{c_{12}b_{24}}{1 - c_{12}c_{21}}x_4.$$

В новых обозначениях коэффициентов α и δ второе приведенное уравнение такое:

$$\hat{y}_1 = \alpha_1 + \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3 + \delta_{14}x_4. \qquad (10.4)$$

Любое приведенное уравнение может быть решено исходя из данных об экзогенных признаках x_j по совокупности единиц (например, предприятий), т.е. могут быть вычислены такие значения коэффициентов регрессии δ_j, при которых согласно МНК достигается минимум суммы квадратов отклонений расчетных значений эндогенной переменной $\hat{y}_i$ от фактических ее значений y_i. Однако не всегда возможно перейти от коэффициентов приведенного уравнения к коэффициентам структурного уравнения. Для того чтобы такой переход, другими словами, расчет коэффициентов a_j, b_{ji} из коэффициентов α_j и δ_{ji}, был возможен и однозначен, требуется соблюдение условия, называемого *условием точной идентификации уравнений*. Это условие может быть выражено в разных формах. Самое простое выражение условия точной идентификации такое: в приведенном уравнении должно быть то же число параметров, что и в структурном. Условие идентифи-

кации можно выразить, и не используя приведенную форму уравнений, так: в правой части структурного уравнения должно отсутствовать столько же экзогенных переменных, входящих в структурные уравнения эндогенных переменных, входящих в правую часть данного структурного уравнения, сколько входит в нее эндогенных переменных.

В нашем примере, исходя из первой формулировки, имеем в каждом приведенном уравнении пять параметров, включая свободные члены. В структурных уравнениях (10.2) было тоже по пять параметров, т.е. условие точной идентификации соблюдено. В соответствии со второй формулировкой в правой части каждого из структурных уравнений отсутствует по одной экзогенной переменной, входящей в уравнение эндогенной переменной, которая входит в эту правую часть: в первом уравнении нет x_4, входящего в уравнение $\hat{y}_2$, а во втором нет x_2, входящего в уравнение $\hat{y}_1$. Число отсутствующих экзогенных переменных равно числу входящих в правые части структурных уравнений эндогенных переменных — условие точной идентификации соблюдено.

Если в правую часть структурных уравнений входят все экзогенные переменные, имеющиеся в уравнениях других эндогенных переменных, и еще эта (эти) эндогенные переменные, то в структурных уравнениях будет больше параметров, чем в приведенных. Тогда из меньшего числа найденных коэффициентов окажется невозможно определить большее число коэффициентов структурного уравнения. Система решения не имеет и называется *неидентифицируемой*. То же будет и при отсутствии в правой части структурных уравнений меньшего числа экзогенных переменных, чем там присутствует эндогенных. Положение неидентификации аналогично неразрешимости системы, включающей меньше уравнений, чем в них включено неизвестных величин.

Аналогично и обратное положение: если число уравнений больше, чем число входящих в них неизвестных, то имеется множество возможных решений и возникает проблема выбора одного из них. Если в нашей системе уравнений отсутствует в каждом из них или в одном больше экзогенных переменных, чем в правой части имеется эндогенных переменных, то в приведенных уравнениях окажется больше параметров, чем в структурных уравнениях. Однозначного решения

(перехода) система не имеет. Такая система уравнений называется *сверхидентифицируемой*.

10.4. Косвенный метод наименьших квадратов

Рассмотрим прежде всего методику решения точно идентифицируемой системы, а затем — сверхидентифицируемой системы. Метод решения точно идентифицируемой системы уравнений называется *косвенным методом наименьших квадратов (КМНК)*, так как МНК применяется не прямо к структурным уравнениям, а к приведенным. Полученные значения параметров приведенных уравнений зависят только от входящих в приведенные уравнения экзогенных переменных и не содержат искажающего влияния других факторов на вариацию эндогенных переменных. При алгебраическом преобразовании параметров приведенных уравнений в параметры структурных уровней, естественно, никакие посторонние факторы на результат не влияют. Следовательно, при КМНК мы получим неискаженные, т.е. состоятельные и несмещенные, значения параметров структурных уравнений.

Пример. Рассмотрим систему взаимосвязанных регрессионных уравнений производительности и оплаты труда. В реальном исследовании, как уже сказано в главе 9, для построения надежных уравнений регрессии («моделей») необходима достаточно большая и желательно однородная совокупность. В учебных целях рассмотрим небольшую выборку и простейшую систему уравнений, описывающих связи между следующими переменными:

y_1 — показатель производительности труда, тыс. руб. на 1 работника в месяц;

y_2 — средняя месячная заработная плата работников, тыс. руб. на 1 работника в месяц;

x_1 — энерговооруженность работника, киловатт на 1 работника;

x_2 — средний разряд работы, выполняемой на предприятии.

Структурные уравнения:

$$\hat{y}_1 = a_1 + b_{11}x_1 + c_{12}y_2$$

$$\hat{y}_2 = a_2 + b_{22}x_2 + c_{21}y_1.$$

Таблица 10.1

Точно идентифицируемая система

Предприятие	x_1	x_2	y_1	y_2	Расчетные значения	
					$\hat{y}_1$	$\hat{y}_2$
1	12	4,7	24	0,9	21	0,7
2	12	4,9	18	1,1	22	1,4
3	15	5,1	26	1,8	25	1,7
4	17	4,6	21	0,7	27	1,3
5	18	3,9	33	0,7	27	0,6
6	21	4,7	31	1,8	31	1,5
7	21	4,7	28	1,9	31	1,5
8	22	4,9	37	1,8	32	1,8
9	22	5,1	34	2,1	39	2,0
10	25	5,3	39	2,0	35	2,3
11	27	4,2	32	0,8	37	1,3
12	28	4,8	37	2,4	39	1,9
Σ	240	56,9	360	18,0	366	18,0

В каждом из них в правой части присутствует одна эндогенная переменная и отсутствует по одной экзогенной переменной, влияющей на эндогенную. Условие идентификации соблюдено, и к системе следует применить КМНК. Данные о показателях по 12 предприятиям приведены в табл. 10.1. В ней также содержатся расчетные значения эндогенных переменных, вычисленные по приведенным уравнениям, с округлением их до той же степени точности, как и исходные данные.

Приведенные уравнения для рассматриваемой системы имеют вид:

$$\hat{y}_1 = \alpha_1 + \delta_{11}x_1 + \delta_{12}x_2,$$

$$\hat{y}_2 = \alpha_2 + \delta_{21}x_1 + \delta_{22}x_2,$$

где $\alpha_1 = \dfrac{a_1 + c_{12}a_2}{1 - c_{12}c_{21}}$; $\delta_{11} = \dfrac{b_{11}}{1 - c_{12}c_{21}}$; $\delta_{12} = \dfrac{b_{22}c_{12}}{1 - c_{12}c_{21}}$;

$\alpha_2 = \dfrac{a_2 + c_{21}a_1}{1 - c_{12}c_{21}}$; $\delta_{21} = \dfrac{b_{11}c_{21}}{1 - c_{12}c_{21}}$; $\delta_{22} = \dfrac{b_{22}}{1 - c_{12}c_{21}}$.

В результате решения (обычным МНК) приведенных уравнений получаем числовые значения их параметров:

$$\hat{y}_1 = 6{,}175 + 1{,}009x_1 + 0{,}7758x_2; \quad R^2 = 0{,}7306; \quad R^2_{\text{корр}} = 0{,}630;$$

$$\hat{y}_2 = -4{,}142 + 0{,}03889x_1 + 1{,}035x_2; \quad R^2 = 0{,}657; \quad R^2_{\text{корр}} = 0{,}528.$$

Прежде чем преобразовать приведенные уравнения в структурные, следует убедиться в надежности приведенных уравнений, проверив эту надежность либо по *F*-критерию, либо вычислив среднюю ошибку коэффициента детерминации (см. гл. 9) и *t*-критерий Стьюдента. В компьютерных программах обычно выводится на дисплей таблица дисперсионного анализа (ANOVA).

Для табл. 10.2 критическое значение *F* при уровне значимости 0,05, *d.f.* = 2 и *d.f.* = 9 степенях свободы составляет 4,26, а для уровня значимости 0,01 оно составляет 8,02. Таким образом, вероятность надежности уравнения для $\hat{y}_1$ превышает 0,99, т.е. уравнение имеет высокую надежность.

Значение *F*-критерия в табл. 10.3 показывает, что и второе приведенное уравнение имеет высокую надежность — около 0,99. Если же полученные приведенные уравнения оказались

Таблица 10.2

Дисперсионный анализ для уравнения $\hat{y}_1$

Источник вариации	Сумма квадратов	Степени свободы	Дисперсия на одну степень свободы	*F*-критерий
Объясненная	358	2	179	
Остаточная	132	9	14,4	12,4
Общая	490	11		

Таблица 10.3

Дисперсионный анализ для уравнения $\hat{y}_2$

Источник вариации	Сумма квадратов	Степени свободы	Дисперсия на одну степень свободы	*F*-критерий
Объясненная	2,72	2	1,36	
Остаточная	1,42	9	0,158	8,61
Общая	4,14	11		

бы ненадежными, то нет смысла преобразовывать их параметры в параметры структурных уравнений, которые оказались бы столь же ненадежными.

Проведем преобразование приведенных уравнений в структурные, используя вышеприведенные формулы для коэффициентов α и δ.

1. Разделим δ_{21} на δ_{11}, тогда $\frac{b_{11}c_{21}}{1-c_{12}c_{21}} \div \frac{b_{11}}{1-c_{12}c_{21}} = c_{21}$, таким образом, $c_{21} = \frac{0{,}03889}{1{,}009} = 0{,}03854$.

2. Разделим δ_{12} на δ_{22}, тогда $\frac{b_{22}c_{12}}{1-c_{12}c_{21}} \div \frac{b_{22}}{1-c_{12}c_{21}} = c_{12}$, таким образом, $c_{12} = \frac{0{,}7758}{1{,}035} = 0{,}7496$.

3. Находим знаменатель коэффициентов.

$$1 - c_{21}c_{12} = 1 - 0{,}03854 \cdot 0{,}7496 = 0{,}97115;$$

4. Вычисляем коэффициент b_{11}:

$$b_{11} = \delta_{11} \cdot 0{,}97115 = 1{,}009 \cdot 0{,}97115 = 0{,}98.$$

5. Вычисляем коэффициент b_{22}:

$$b_{22} = \delta_{22} \cdot 0{,}97115 = 1{,}035 \cdot 0{,}97115 = 1{,}0057.$$

6. Подставляем c_{12}; c_{21} и знаменатель в выражения для α_1 и α_2, получаем систему уравнений:

$$6{,}175 \cdot 0{,}97115 = a_1 + 0{,}7496a_2;$$

$$-4{,}14 \cdot 0{,}97115 = a_2 + 0{,}03854a_1;$$

$$\{a_1 + 0{,}7496a_2 = 5{,}997;$$

$$\{a_1 + (1 \div 0{,}03854)a_2 = \frac{-4{,}14 \cdot 0{,}97115}{0{,}03854} = -104{,}32.$$

Решая эту систему уравнений, получим:

$$a_1 = 9{,}278;\ a_2 = -4{,}377.$$

Итак, структурные уравнения получили числовую оценку своих коэффициентов и могут быть записаны как:

$$\hat{y}_1 = 9{,}279 + 0{,}98x_1 + 0{,}7496\hat{y}_2;$$

$$\hat{y}_2 = -4{,}377 + 1{,}0057x_2 + 0{,}03854\hat{y}_1.$$

Напомним, что стоящие в правой части структурных уравнений значения эндогенных переменных — это не их исходные значения y_{1i}; y_{2i} из табл. 10.1, а значения, рассчитанные по приведенным уравнениям. Иначе весь смысл КМНК утрачивается и оценки окажутся смещенными. Отсюда следует, что точно идентифицируемая система уравнений вместо КМНК может быть решена и *двойным методом наименьших квадратов (ДМНК)*, а именно, вычислив по решенным приведенным уравнениям значения эндогенных переменных, подставим их и входящие в структурные уравнения экзогенные переменные (их первичные значения) и решаем еще раз обычным МНК сами структурные уравнения, получая их коэффициенты. Заметим, что из-за округления значений эндогенных переменных $\hat{y}_1$ и $\hat{y}_2$ в табл. 10.1 результаты решений методами КМНК и ДМНК абсолютно точно не совпадут.

10.5. Двойной метод наименьших квадратов

Если изучаемая система уравнений является сверхидентифицируемой, решить приведенные уравнения можно, но преобразовать полученные параметры в параметры структурных уравнений однозначно нельзя, так как структурные уравнения содержат меньше коэффициентов, чем приведенные. Следовательно, КМНК не позволяет решить сверхидентифицируемую систему, и нужно идти путем исключения влияния неучтенных факторов на эндогенные переменные, т.е. применить двойной метод наименьших квадратов. Алгоритм ДМНК состоит из следующих последовательных «шагов».

1. Структурные уравнения преобразовывают в приведенные.
2. Приведенные уравнения решаются с помощью МНК.
3. Проверяется надежность уравнений по F-критерию.

4. Если уравнения надежны, по ним вычисляются расчетные значения эндогенных переменных для каждой единицы совокупности.
5. Эти расчетные значения эндогенных переменных, находящихся в правой части структурных уравнений, и соответствующие значения экзогенных переменных используются для решения структурных уравнений с помощью МНК.
6. Вновь проверяется надежность полученных решений. Эта проверка необходима, так как при ДМНК решенные структурные уравнения качественно отличны от приведенных уравнений, в том числе имеют другое число степеней свободы вариации, поэтому надежность приведенных уравнений еще не гарантирует надежности решения структурных уравнений.

Следует предостеречь изучающих данную тему от возможной ошибки: при втором МНК-решении расчетные значения эндогенных переменных, полученные при решении приведенных уравнений, подставляются только в правую часть каждого структурного уравнения, а в его левой части, разумеется, должны оставаться фактические значения определяемой эндогенной переменной для каждой единицы совокупности.

Структурные уравнения:

$$\hat{y}_1 = a_1 + b_{11}x_1 + c_{12}y_2;$$
$$\hat{y}_2 = a_2 + b_{22}x_2 + b_{23}x_3 + c_{21}y_1,$$

где y_1 — производительность труда, руб./ч;
y_2 — среднечасовая оплата труда, руб./ч;
x_1 — энерговооруженность работников, кВт/чел.;
x_2 — средний разряд выполненных работ;
x_3 — коэффициент рентабельности предприятия, %.

В первом уравнении отсутствуют две экзогенные переменные, входящие в уравнение y_2, а эндогенная переменная в его правой части только одна — уравнение сверхидентифицируемое.

Построим приведенные уравнения:

$$\hat{y}_1 = \alpha_1 + \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3;$$
$$\hat{y}_2 = \alpha_2 + \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3.$$

В приведенных уравнениях восемь коэффициентов, а в структурных всего семь — следовательно, КМНК неприменим: однозначно преобразовать восемь коэффициентов в семь нельзя. Нет смысла приводить выражения α_1, α_2 и всех коэффициентов δ через коэффициенты структурных уравнений.

В результате решения получены приведенные уравнения:

$$\hat{y}_1 = -8{,}163 + 0{,}46807x_1 + 2{,}9309x_2 + 0{,}04659x_3;$$
$$R_1^2 = 0{,}7209; \qquad R^2_{\text{корр}} = 0{,}616;$$
$$\hat{y}_2 = -7{,}805 + 0{,}03694x_1 + 2{,}7304x_2 + 0{,}02057x_3;$$
$$R_2^2 = 0{,}715; \qquad R^2_{\text{корр}} = 0{,}608.$$

В результате дисперсионного анализа получены значения F-критерия:

для уравнения $\hat{y}_1$ $\qquad F_1 = 6{,}89;$

для уравнения $\hat{y}_2$ $\qquad F_2 = 6{,}68.$

Табличное значение F-критерия для уровня значимости 0,05 и при 3 и 8 степенях свободы вариации равно 4,46. Таким образом, оба приведенных уравнения достаточно надежны.

Следующий этап решения — вычисление для каждой единицы совокупности расчетных значений эндогенных переменных $\hat{y}_{1i}$, $\hat{y}_{2i}$. Эти значения приведены в 7-й и 8-й графах табл. 10.4. Записывая эти значения для последующих действий, необходимо обеспечить запас точности — не менее 4—5 значащих цифр. Затем эти расчетные значения, а также фактические значения имеющихся в структурных уравнениях экзогенных переменных вводятся в качестве факторов («регрессоров»), а фактические значения эндогенных переменных, стоящих в левых частях структурных уравнений, — в качестве зависимых переменных, и решаются структурные уравнения на ЭВМ, т.е. осуществляется вторичное МНК-решение.

Последний этап: записываются параметры структурных уравнений и проверяется заново их надежность по F-критерию. В нашем примере структурные уравнения получили вид:

$$\hat{\hat{y}}_1 = -0{,}0529 + 0{,}4286x_1 + 1{,}182\hat{y}_2; \qquad R^2_{\text{корр}} = 0{,}655;$$
$$\hat{\hat{y}}_2 = -6{,}963 + 2{,}599x_2 + 0{,}02028x_3 + 0{,}04144\hat{y}_1; \qquad R^2_{\text{корр}} = 0{,}570.$$

Для уравнения $\hat{\hat{y}}_1$ $F = 11{,}42$, а табличное значение F-критерия при 2 и 9 степенях свободы и уровне значимости 0,05 равно 4,26, а при уровне значимости 0,01 равно 8,02. Следовательно, надежность установления связи превышает $(1 - -$ 0,01), т.е. превышает 0,99. Для уравнения $\hat{\hat{y}}_2$ $F = 5{,}87$, а табличное значение при 3 и 8 степенях свободы и уровне значимости 0,05 равно 4,46. Надежность установления связи для этого уравнения превышает $(1 - 0{,}05)$, т.е. больше 0,95.

При обозначении переменных в структурных уравнениях во избежание путаницы стоящие в правой части эндогенные переменные обозначены с одной чертой (или «домиком»), т.е. как $\hat{y}_1$ и $\hat{y}_2$. Это означает, что они являются расчетными значениями после *первого* применения МНК к приведенным уравнениям. А те же эндогенные y_1 и y_2, стоящие в левой части решений структурных уравнений, обозначены с двумя чер-

Таблица 10.4

Сверхидентифицируемая система уравнений

Первичные данные						Расчетные значения			
						по приведенным		по структурным	
№ единицы совокупности	x_1	x_2	x_3	y_1	y_2	$\hat{y}_1$	$\hat{y}_2$	$\hat{\hat{y}}_1$	$\hat{\hat{y}}_2$
А	1	2	3	4	5	6	7	8	9
1	21	4,0	17	12	3,5	14,182	4,241	14,06	4,367
2	17	4,6	7	12	5,6	13,602	5,526	13,87	5,700
3	30	4,3	8	14	4,2	18,855	5,208	19,06	5,158
4	22	5,0	10	16	6,8	17,255	6,865	17,59	6,952
5	20	4,1	10	17	4,9	13,961	4,456	13,89	4,598
6	31	4,7	45	18	5,8	22,219	7,096	21,72	7,088
7	32	4,2	52	23	6,8	21,548	5,911	20,75	5,902
8	34	4,2	10	24	5,4	20,527	5,124	20,68	5,008
9	28	4,8	21	24	7,4	19,990	6,766	20,05	6,769
10	42	4,3	8	25	6,3	24,472	5,651	24,73	5,391
11	32	5,0	24	26	8,2	22,588	7,512	22,64	7,457
12	51	4,8	22	29	7,1	30,802	7,636	30,93	7,610
Σ	360	54	240	240	72	240,001	71,992	239,97	72,000

точками («домиками»). Это означает, что они являются расчетными значениями после *двойного* применения МНК. Эти значения приведены в последних графах табл. 10.4. Как видим, они не совпадают со значениями, полученными по приведенным уравнениям. Ведь состав факторов в структурных и в приведенных уравнениях неодинаков. Заметим, что об этом обстоятельстве, очень важном, как правило, не упоминается.

РЕЗЮМЕ

Уравнение множественной регрессии описывает связь между независимыми переменными («входами») и зависимой переменной («выходом»). Оно не раскрывает механизма связи между всеми переменными и в этом смысле соответствует модели «черного ящика». Этим определяется важность построения системы уравнений регрессии, соответствующих всей системе связей между переменными.

Для каждой конкретной задачи признаки, подлежащие определению, называются эндогенными, а переменные, считающиеся для данной задачи заданными (известными), — экзогенными.

Если каждая из эндогенных переменных является только зависимой, то соответствующая система уравнений называется рекуррентной (или рекурсивной).

Метод наименьших квадратов обеспечивает получение несмещенных оценок параметров, если корреляция между уточненными объясняющими переменными («ошибками») отсутствует.

Система уравнений, соответствующая структуре связей, называется системой структурных уравнений.

Уравнение, которое в правой части не содержит эндогенных переменных, называется приведенным.

Для однозначного перехода от коэффициента приведенных уравнений к коэффициентам структурных уравнений требуется выполнение условия точной идентификации.

Самое простое выражение точной идентификации состоит в том, что в приведенном уравнении должно быть то же число параметров, что и в структурном. Условие идентификации можно сформулировать так: в правой части структур-

ного уравнения должно отсутствовать столько же экзогенных переменных, сколько входит в нее эндогенных переменных.

Если в правую часть структурных уравнений входят все экзогенные переменные, имеющиеся в уравнениях других экзогенных переменных, то система не имеет решения и называется неидентифицируемой. Если в каждом из уравнений системы или в одном из них больше экзогенных переменных, чем эндогенных переменных в правой части уравнения, то такая система называется сверхиндентифицируемой.

Оценка параметров идентифицируемой системы проводится косвенным методом наименьших квадратов (КМНК) или двойным методом наименьших квадратов (ДМНК).

Оценка параметров сверхидентифицируемой системы проводится ДМНК.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика и основы эконометрики: Учебник. 2-е изд. — М.: ЮНИТИ, 2001.
2. *Бородич С. А.* Эконометрика. Учеб. пособие. — Минск: Новое знание, 2001.
3. *Джонстон Дж.* Эконометрические методы. — М.: Статистика, 1980.
4. *Магнус Я. Р., Катышев П. К., Пересецкий А. А.* Эконометрика: Начальный курс. 2-е изд. — М.: Дело, 2000.
5. *Тинтнер Г.* Введение в эконометрию. — М.: Финансы и статистика, 1965.
6. *Фишер Ф.* Проблема идентификации в эконометрии. — М.: Статистика, 1978.
7. *Эконометрика*: Учебник / Под ред. И. И. Елисеевой. — М.: Финансы и статистика, 2002.

Глава
11

СТАТИСТИЧЕСКИЙ АНАЛИЗ НЕКОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ

11.1. Зависимость методов измерений связей от уровня измерения переменных

Рассмотренные методы корреляционного и регрессионного анализов разработаны для переменных, измеренных на *интервальной шкале* или *шкале отношений* (см. гл.1). Интервальные шкалы могут быть построены лишь для количественных признаков, позволяющих не только упорядочить объекты, но и рассчитать величину отличия (интервал) одной степени появления признака от другой. Примерами интервальных шкал могут служить шкалы измерения большинства экономических характеристик.

В случаях, когда можно указать абсолютный нуль на шкале, мы имеем шкалу отношений. По такой шкале можно, сопоставляя переменные, заключить, что одно значение больше (меньше) другого в два раза и т.п. По шкале отношений можно измерять такие характеристики, как стаж работы, заработная плата, результаты голосования, потребление природного газа, окупаемость инвестиций и т.п.

Для переменных, измеренных по интервальной шкале и шкале отношений, можно использовать все арифметические действия, включая извлечение корней, возведение в степень и логарифмирование. Поэтому можно измерять связь между такими переменными на основе ковариации $\sum_{(i)}(x_i - \bar{x})(y_i - \bar{y})$ и разложения дисперсии, а также построения неких функций

(уравнений регрессии) , описывающих среднее изменение зависимой переменной при изменении одной или нескольких независимых переменных : $\hat{y} = \hat{f}(x_1, x_2, \ldots, x_k)$.

Иное дело, когда нам приходится обобщать и анализировать те свойства объектов, которые измерены на более низких уровнях: на *порядковой (ординальной) шкале* или на *номинальной шкале*. Как показано в гл.1, измерение на номинальной шкале — это просто указание градации признака x для данного объекта. Градации номинальной шкалы должны быть образованы так, чтобы различия внутри классов были малы, а между классами — сравнительно велики, при этом классы не должны перекрывать друг друга.

Градации на номинальной шкале могут быть как некоторыми высказываниями, так и числами. Например, для характеристики клиентов компьютерной фирмы можно использовать перечень категорий клиентов:

- государственные организации;
- муниципальные организации;
- частные фирмы;
- индивидуальные пользователи,

а можно присвоить цифровые метки каждой категории: 1, 2, 3, 4. Но следует помнить, что числа на этой шкале играют роль ярлыков и к ним неприменимы обычные правила арифметики. Номинальная шкала обладает только свойствами *симметричности* и *транзитивности*. Симметричность означает, что отношения, существующие между градациями x_2 и x_1, соответствуют отношениям между x_1 и x_2. Транзитивность означает, что если $x_1 = x_2$ и $x_2 = x_3$, то $x_1 = x_3$.

Порядковая шкала — следующий, более высокий уровень измерения. Измерение по порядковой шкале означает, что мы не только получаем знание о том, к какой категории по данному признаку принадлежит объект, но и о том, в каком отношении он находится с другими объектами, принадлежащими к иным категориям по данному признаку. Градации порядковой шкалы упорядочены между собой некоторым асимметричным (в отличие от номинальной шкалы) образом, т.е. если имеется $x_2 > x_1$, то несправедливо отношение $x_1 > x_2$. Свойство транзитивности выполняется, если $x_3 > x_2$ и $x_2 > x_1$, то $x_3 > x_1$. Эти свойства позволяют расположить x_1, x_2, x_3... по возрастающей или убывающей степени выраженности признака. Необходимо помнить, что при измерении по порядковой шкале мы не получаем информации *о величине различий* между объектами по данному признаку, мы только устанавливаем некоторый порядок следования объектов, например, что x_3 больше x_1, но мы не можем сказать, что различия между x_1 и x_2 больше, чем между x_2 и x_3. Так что к градациям порядковой шкалы неприменимы обычные арифметические действия.

Номинальные и порядковые переменные традиционно входят в социологическую информацию, собираемую в форме опросов, интервью, заполнения вопросников. Однако все чаще такого рода переменные появляются в данных, собираемых в экономических исследованиях. Причина этого прежде всего в стремлении отразить человеческий фактор в бизнесе. Во всем мире и в России используется система бизнес-обследований на основе опросов руководителей организаций, когда всем участникам прошлого опроса рассылают новую анкету и одновременно представляют результаты предыдущего опроса, что повышает мотивацию опрашиваемых к точности и оперативности ответов. Обычно респондентам предлагается оценить фактическое состояние и ожидаемое изменение основных показателей хозяйственной деятельности в рамках альтернатив: «увеличение—уменьшение», «улучшение—ухудшение», «сохранение на прежнем уровне». К такого же рода информации относятся ответы на вопросы, связанные с выявлением факторов, ограничивающих деятельность организации, и другие качественные вопросы. Понятно, что такого рода данные могут быть получены только в результате опросов. Собираемые данные лишь отражают оценки и ожидания менеджеров, которые склонны думать, что ни плохие, ни хорошие периоды в деятельности фирмы не могут продолжаться вечно. Обобщение таких данных позволит сделать вывод относительно краткосрочных перспектив развития; они полезны при разработке экономических индикаторов.

Приведем примеры вопросов, задаваемых в бизнес-обследованиях.

Вопрос	Варианты ответа
1. Объем производства (по отношению к предыдущему периоду)	1. Выше 2. Без изменений 3. Ниже

Вопрос	Варианты ответа
2. Спрос на продукцию	1. Выше нормального уровня 2. Нормальный 3. Ниже нормального уровня («нормальный» — с точки зрения субъективных представлений респондента)
3. Численность занятых на предприятии на момент обследования (чел.)	1. Не более 100 2. 101—1000 3. 1001—5000 4. 5001—10 000 5. Свыше 10 000

Сопутствующим фактором расширения опросов и оперирования неколичественными данными является *коммерческая тайна*. Вам могут не назвать конкретную сумму прибыли (убытка) за период, но скажут, как изменилось финансовое положение предприятия: улучшилось, ухудшилось, осталось прежним. Часто ответы на подобные вопросы представляются в виде *шкалы Ликерта*, включающей крайние позиции (позитивную и негативную), промежуточные и нулевую точку. Например, для ответа на вопрос: «Как вы оцениваете перспективы развития вашей организации?» могут быть предусмотрены следующие варианты:

- очень хорошие;
- скорее хорошие, чем плохие;
- трудно сказать;
- скорее плохие, чем хорошие;
- очень плохие.

Такого рода данные можно упорядочивать, можно приписать цифровые метки каждому варианту ответа, например: 1; 0,5; 0; −0,5; −1. Но это вовсе не означает, что перспективы развития одних предприятий вдвое лучше или хуже перспектив других предприятий, так как эти данные относятся к порядковым.

Порядковые данные привлекают все больше внимания в связи с построением *рейтингов* коммерческих банков, высших учебных заведений, торговых и промышленных организаций администраторов, политических деятелей, артистов, спортсменов. Рейтинг — по сути, порядковая переменная.

Еще одним фактором расширения сферы переменных, измеренных по номинальной или порядковой шкале, служит повышение внимания к социальным проблемам (выявление удовлетворенности людей разными сторонами жизни — работой, отношениями в семье, их мотивацией, жизненными планами и т.д.).

Обработка и анализ данных, измеренных по разным шкалам, должны проводиться особыми методами. Зависимость методов измерения связей от уровня измерения переменных можно проследить по табл. 11.1.

Из табл. 11.1 следует важный вывод: меры связей, разработанные для переменных более низкого уровня измерения, могут использоваться для измерения связей между переменными более высокого уровня измерения. Говоря иными словами, меры, которые разработаны для номинальных переменных, могут быть рассчитаны и для измерения связей между порядковыми или интервальными переменными. Заметим, что при этом происходит потеря информации, поскольку вместо того, чтобы зафиксировать конкретную сумму дохода домохозяйств, скажем 155 тыс. руб. в год и т.д., мы создаем некие градации величин дохода: до 80 тыс. руб.; 80—100 тыс. руб.; 100—150;150—200; 200—250 тыс. руб. и т.д. и подсчитываем число домохозяйств с данной суммой годового дохода. Тем самым мы теряем часть информации о вариации годового дохода.

На основе значений порядковых переменных единицам (объектам) приписываются *ранги*, и связь между порядковыми переменными измеряется на основе совпадения (или несовпадения) рангов по разным признакам. Таким образом, измерение корреляции между порядковыми переменными в неявном виде учитывает значение переменной.

При измерении связи между номинальными переменными значения переменных не участвуют в расчетах: меры связей основаны на частоте совместного появления определенной i-й категории (x_i) одной переменной x и определенной j-й категории (y_j) другой переменной y. Измерение связей между номинальными переменными основывается на клеточных частотах таблицы сопряженности n_{ij} (см. подразд. 8.3).

Таблица 11.1

Меры тесноты связей

Шкала измерения переменной x	Шкала измерения переменной y		
	номинальная	порядковая	интервальная или шкала отношений
Номинальная	Теоретико-информационные меры связей. Коэффициенты взаимной сопряженности и другие показатели, основанные на таблицах сопряженности	Теоретико-информационные меры связей. Коэффициенты взаимной сопряженности и другие показатели, основанные на таблицах сопряженности	Те же меры связей, а также меры связей, основанные на разложении дисперсии зависимой переменной (эмпирическое корреляционное отношение)
Порядковая	Теоретико-информационные меры связей. Коэффициенты взаимной сопряженности и другие показатели, основанные на таблицах сопряженности	Коэффициенты ранговой корреляции, коэффициент конкордации	Те же меры связей, а также меры связей, основанные на разложении дисперсии зависимой переменной (эмпирическое корреляционное отношение)
Интервальная или шкала отношений	Теоретико-информационные меры связей. Коэффициенты взаимной сопряженности и другие показатели, основанные на таблицах сопряженности	Коэффициенты ранговой корреляции, коэффициент конкордации	Те же меры связей. Коэффициенты парной, частной, множественной корреляции. Индексы корреляции

11.2. Измерение связи между двумя дихотомическими переменными

Для измерения связи между двумя дихотомическими переменными (т.е. признаками, каждый из которых принимает два значения) данные представляются в виде таблицы сопряженности 2 × 2 (ее называют также четырехпольной таблицей). Например, изучается связь между активностью работы в профсоюзе и уровнем заработной платы (табл. 11.2).

В табл. 11.2 показано, как распределились по категориям 100 работников, по которым были получены данные о зарабо-

Таблица 11.2

Активность в профсоюзе и уровень заработной платы

Проявление активности	Уровень заработной платы		Итого
	высокий	низкий	
Высокая	45	5	50
Низкая	15	35	50
Итого	60	40	100

ботной плате и работе в профсоюзе. Очевидно, что эти переменные связаны: появление лиц с сочетанием высокой активности (или неактивности) в профсоюзе и высоким (низким) уровнем заработной платы не является равновероятным. Среди тех, кто активно работает в профсоюзе, вероятность встретить высокооплачиваемых работников гораздо выше, чем среди тех, кто не отличается активностью. Для таких таблиц разработаны специальные меры связей. К ним относятся: коэффициент ассоциации, коэффициент контингенции.

Коэффициент ассоциации Q предложен английским статистиком Дж. Э. Юлом (1871—1951):

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}},$$

где n_{11} — число единиц, имеющих значения x_1 и y_1;
n_{22} — число единиц, имеющих значения x_2 и y_2;
n_{12} — число единиц, имеющих значения x_1 и y_2;
n_{21} — число единиц, имеющих значения x_2 и y_1.

Общий вид таблицы сопряженности 2 × 2 представлен в табл. 11.3.

Коэффициент ассоциации Q принимает значения в интервале [0,1]: 0 — отсутствие связи, 1 — полная связь. Вычислим значение Q по данным табл. 11.2:

$$Q = \frac{45 \cdot 35 - 15 \cdot 5}{45 \cdot 35 + 15 \cdot 5} = \frac{1500}{1650} = 0{,}909,$$

т.е. связь между изучаемыми признаками очень тесная.

Таблица 11.3

Таблица 2 × 2

Переменная x	Переменная y		Итого
	y_1	y_2	
x_1	n_{11}	n_{12}	$n_{1.} = n_{11} + n_{12}$
x_2	n_{21}	n_{22}	$n_{2.} = n_{21} + n_{22}$
Итого	$n_{.1} = (n_{11} + n_{21})$	$n_{.2} = (n_{12} + n_{22})$	$n = n_{1.} + n_{2.} = n_{.1} + n_{.2} = n_{11} + n_{12} + n_{21} + n_{22}$

В случае отсутствия связи между активностью и заработной платой мы бы имели в каждой клетке табл. 11.2 по 25 человек, и тогда Q был бы равен:

$$Q = \frac{25 \cdot 25 - 25 \cdot 25}{25 \cdot 25 + 25 \cdot 25} = \frac{0}{1250} = 0.$$

Мера связи Юла основана на сравнении вероятности появления взаимно совместимых (гомогенных) и взаимно несовместимых (гетерогенных) пар значений. Взаимно совместимыми в нашем примере являются : «высокая активность — высокая заработная плата», «низкая активность — низкая заработная плата»; взаимно несовместимыми являются: «низкая активность — высокая заработная плата», «высокая активность — низкая заработная плата».

Коэффициент ассоциации принимает значение «1», если хотя бы одна из клеток таблицы 2 × 2 равна нулю (см. например, табл. 11.4 и 11.5).

Таблица 11.4

Случай полной связи

	y_1	y_2	Итого
x_1	—	50	50
x_2	50	—	50
Итого	50	50	100

Таблица 11.5

Случай неполной связи

	y_1	y_2	Итого
x_1	25	—	25
x_2	30	45	75
Итого	55	45	100

Для табл. 11.4: $Q = -1$, случай полной связи.

Для табл. 11.5: $Q = 1$, хотя связь между x и y далеко не полная.

Эта особенность коэффициента ассоциации снижает его значение и показывает, насколько важно соблюдать осторожность при интерпретации результатов измерения связи.

Более достоверное измерение связи обеспечивает *коэффициент контингенции,* обозначаемый либо Ф («фи»), либо $K_{\text{контингенции}}$:

$$\Phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11} + n_{12})(n_{22} + n_{21})(n_{11} + n_{21})(n_{12} + n_{22})}} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{.1}n_{.2}n_{1.}n_{2.}}}, \quad (11.1)$$

где $n_{1.}$ и $n_{2.}$ — итоги по строкам таблицы;

$n_{1.}$ и $n_{2.}$ — итоги по столбцам;

$-1 \leqslant \Phi \leqslant 1$.

По данным табл. 11.5, для которой было найдено $Q = 1$,

$$\Phi = \frac{25 \cdot 45 - 30 \cdot 0}{\sqrt{(25 + 0)(30 + 45)(25 + 30)(0 + 45)}} = \frac{1125}{2154} = 0{,}52.$$

Для табл. 11.4, где связь полная, коэффициент контингенции также равен 1:

$$\Phi = \frac{0 \cdot 0 - 50 \cdot 50}{\sqrt{(0 + 50)(50 + 0)(0 + 50)(50 + 0)}} = \frac{-2500}{2500} = -1.$$

По данным табл. 11.2, для которых коэффициент ассоциации составил $Q = 0{,}909$, коэффициент контингенции равен:

$$\Phi = \frac{45 \cdot 35 - 15 \cdot 5}{\sqrt{(45+5)(15+35)(45+15)(5+35)}} = \frac{1500}{2449} = 0{,}612,$$

т.е. значение Ф существенно ниже величины Q. Приведенные примеры подтверждают, что коэффициент контингенции является более достоверной мерой связи между дихотомическими переменными. Можно показать, что формула Ф получена на основе формулы коэффициента парной корреляции К. Пирсона r. Соответственно свойства коэффициента контингенции такие же, как и у коэффициента корреляции: коэффициент контингенции принимает нулевое значение, если оба произведения в числителе точно уравновешиваются (что крайне маловероятно); он равен -1, если отсутствуют гомогенные сочетания, $n_{11} = 0$ и $n_{22} = 0$.

Важным частным случаем задачи является измерение связи при альтернативной вариации двух признаков, один из которых имеет характер причины, а другой — следствия. Например, при социологическом исследовании 1000 жителей города были поставлены два вопроса: 1. Считаете ли вы, что ваши доходы обеспечивают удовлетворение основных потребностей? 2. Удовлетворяет ли вас деятельность мэра города? Можно предположить, что причиной отрицательного ответа на 2-й вопрос у части населения является неудовлетворенность их потребностей доходами, т.е. имеется связь между ответами на оба вопроса. Для измерения этой связи составляют двухмерное (дихотомическое) распределение ответов 2 × 2, приведенное в табл. 11.6.

Если бы все ответившие «да» на 1-й вопрос отвечали бы «да» на 2-й вопрос, и так же совпадали ответы «нет», то связь была бы предельно тесной, функциональной. Но на самом деле распределение ответов на оба вопроса не совпадает. Большая часть ответивших «да» на 1-й вопрос ответили «да» и на 2-й вопрос, но часть ответила «нет». То же относится к ответившим «да» на 2-й вопрос. Связь есть, но неполная, типа корреляционной, и нужно измерить тесноту этой связи.

Таблица 11.6

Взаимосвязь между ответами на два вопроса социологического обследования

Ответы на 1-й вопрос	Ответы на 2-й вопрос		Итого
	да (a)	нет (b)	
Да (A)	170	80	$\Sigma A = 250$
Нет (B)	230	520	$\Sigma B = 750$
Итого	$\Sigma a = 400$	$\Sigma b = 600$	$n = 1000$

Можно предложить показатель тесноты связи в форме отношения избытка суммы гомогенных сочетаний над их пропорциональной суммой к предельно возможному избытку.

Для этого необходимо вначале вычислить, каковы были бы пропорциональные числа гомогенных сочетаний Aa и Bb? Пропорциональные числа — это доли от общей численности совокупности n, которые были бы получены при полном отсутствии взаимосвязи группировок по двум признакам (ответам на два вопроса), т.е. числа $(\Sigma A \cdot \Sigma a : n)$ и $(\Sigma B \cdot \Sigma b : n)$, составляющие по данным табл. 11.6:

$$Aa = \frac{250 \cdot 400}{1000} = 100 \quad \text{и} \quad Bb = \frac{750 \cdot 600}{1000} = 450.$$

При отсутствии связи на первой диагонали таблицы в сумме было бы: 100 + 450 = 550 единиц совокупности, а на самом деле их: 170 + 520 = 690. Избыток, образовавшийся ввиду прямой связи между ответами, составил: 690 − 550 = 140.

Предельно возможный избыток был бы в том случае, если бы не было гетерогенных сочетаний, т.е. Ab и Ba. Он составляет: 140 + 80 + 230 = 450. Сам же показатель тесноты связи — отношение фактического излишка к предельному: 140 : 450 = = 0,311. Как видим, этот показатель близок к коэффициенту ассоциации, но обладает чрезвычайно логичной и ясной интерпретацией: связь составляет 0,311, или 31,1% предельно возможной функциональной. Данный показатель — аналог не коэффициента корреляции, а коэффициента детерминации. Поэтому правомерно обозначить его как R^2 или η^2. Он имеет вид:

$$\eta^2 = \frac{Aa + bB - |Aa' + Bb'|}{n - (Aa + Bb)}, \qquad (11.2)$$

где

$$Aa' = \frac{\sum A \sum a}{n}; \qquad Bb' = \frac{\sum B \sum b}{n}.$$

Подставляя эти выражения в (11.2), получим:

$$\eta^2 = \frac{Aa + Bb - \frac{\sum A \sum a + \sum B \sum b}{n}}{n - \frac{\sum A \sum a + \sum B \sum b}{n}} =$$

$$= \frac{n(Aa + Bb) - (\sum A \sum a + \sum B \sum b)}{n^2 - (\sum A \sum a + \sum B \sum b)}. \qquad (11.3)^1$$

11.3. Измерение связи по таблицам взаимной сопряженности $m \times p$

В гл. 8 рассмотрено применение непараметрического критерия хи-квадрат для испытания гипотезы о независимости двух переменных. Если нулевая гипотеза об отсутствии связи отклоняется, $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, то необходимо измерить тесноту связи. Само значение критерия хи-квадрат в качестве меры связи не используется, хотя, конечно, большая величина хи-квадрата дает основание надеяться на то, что связь между переменными будет тесной. Но величина хи-квадрата зависит от числа наблюдений n, от распределения наблюдений по клеткам таблицы, т.е. от клеточных частот n_{ij}, а значит, и от числа категорий, выделяемых по одной переменной m и по другой переменной p, т.е. величина критерия хи-квадрат зависит от числа строк и столбцов таблицы. Вдобавок значение хи-квадрата не имеет верхнего предела $0 \leqslant \chi^2 < \infty$. Поэтому измерение связи между категоризованными переменными проводится с помощью специальных мер связи. Для таблиц

[1] Эта мера связи предложена М. Юзбашевым в статье «О новом показателе тесноты связи описательных признаков» // Вестник статистики. — 1986. — № 3. — С. 65—66.

размерности m и p используют прежде всего *коэффициенты взаимной сопряженности*. В эту группу показателей входят коэффициенты взаимной сопряженности К.Пирсона, А.Чупрова, Г.Крамера. Все эти меры основаны на критерии хи-квадрат. Как и все статистические меры связи, коэффициенты взаимной сопряженности принимают значение в интервале [0, 1]. Равенство коэффициента нулю означает отсутствие связи, равенство единице означает полную связь.

Предварительный вывод о тесноте связи можно сделать на основе просмотра таблицы сопряженности: если клеточные частоты главной диагонали заметно больше остальных клеточных частот — связь тесная. Это может быть диагональ из левого верхнего угла таблицы в правый нижний угол или из левого нижнего угла в правый верхний угол таблицы.

К.Пирсон установил, что для нормально распределенных признаков

$$\varphi^2 = \frac{r^2}{1 - r^2},$$

где $\varphi^2 = \chi^2 : n$;

r — коэффициент парной корреляции (линейной).

Отсюда можно получить выражение меры связи для таблицы сопряженности:

$$r = \sqrt{\frac{\varphi^2}{1 - \varphi^2}}.$$

Обозначив эту меру буквой P, получим формулу коэффициента взаимной сопряженности К. Пирсона:

$$P = \sqrt{\frac{\varphi^2}{1 + \varphi^2}}; \qquad (11.4)$$

$$\varphi^2 = \sum_{(i)} \sum_{(j)} \frac{n_{ij}^2}{n_i n_j} - 1, \qquad (11.5)$$

где i — номер категории по признаку x, $i = 1 \ldots m$;

j — номер категории по признаку y, $j = 1 \ldots p$;

$$0 \leqslant P \leqslant 1.$$

Пример. Рассмотрим пример расчета коэффициента взаимной сопряженности. Воспользуемся данными ФРГ, где для новорожденных регистрируется религиозная принадлежность отца и матери. Эти данные публикуются в статистическом ежегоднике ФРГ.

При этом выделены 5 групп по религиозной принадлежности граждан: 1) евангелическая (в России их чаще именуют протестантами); 2) римско-католическая; 3) прочие христиане (включая и православных); 4) другие религии; 5) неверующие или не указавшие религиозную принадлежность (табл. 11.7).

В табл. 11.7 представлена «решетка» 5 × 5, и все ее клетки непусты: встречаются детные браки между лицами любых вероисповеданий. При этом наибольшие числа располагаются вдоль «главной диагонали», т.е. явно преобладают случаи, когда отец и мать ребенка придерживаются одной и той же религии. Такое явление ярче всего просматривается среди лиц «других религий» — магометан, иудеев, буддистов,

Таблица 11.7

Распределение новорожденных в ФРГ по принадлежности отца и матери в

(тыс. чел.)

Религия отца	Религия матери					Итого $n_{i.}$
	евангелическая	римско-католическая	прочие христиане	другие религии	неверующие или не указавшие религиозную принадлежность	
Евангелическая	146,1	57,6	1,1	0,5	8,8	214,1
Римско-католическая	57,3	195,9	1,1	0,7	5,2	260,2
Прочие христиане	1,3	1,4	10,5	0,1	0,3	13,6
Другие религии	1,8	2,0	0,1	62,8	1,1	67,8
Неверующие или не указавшие религиозную принадлежность	29,1	16,1	0,7	0,8	77,7	124,4
Итого $n_{.j}$	235,6	273,0	13,5	64,9	93,1	680,1

Источник. Statistisches Jahrbuch für die BRD. — 1995. — С. 74.

индуистов. Среди них в 92,6% оба родителя придерживаются одной и той же веры. Среди лиц евангелического вероисповедания только 68,2% родительских пар придерживаются одной и той же веры.

Гипотеза о связи заключается в том, что существует предпочтение к заключению брака между лицами одинакового вероисповедания.

Соответствующие этой гипотезе частоты n_{ij} расположены вдоль первой диагонали таблицы (из верхнего левого угла в правый нижний). Они явно превосходят недиагональные частоты. $\chi^2_{\text{факт}} = 1453$, эта величина критерия превосходит его критическое значение как на 5%-ном уровне значимости, так и на 1%-ном уровне значимости:

$\chi^2_{\alpha = 0,05,\ d.f. = 16} = 26,3$; $\chi^2_{\alpha = 0,01,\ d.f. = 16} = 32,0$.

Следовательно, принадлежность отца и матери к одной и той же религии не случайна. Теперь измерим тесноту этой связи с помощью коэффициента взаимной сопряженности К. Пирсона P:

$$\varphi^2 = \frac{146,1^2}{214,1 \cdot 235,6} + \frac{57,6^2}{214,1 \cdot 273} + \ldots + \frac{77,7^2}{124,4 \cdot 93,1} - 1 = 2,1364;$$

$$P = \sqrt{\frac{2,1364}{3,1364}} = 0,825.$$

Связь тесная.

Недостаток коэффициента Пирсона в том, что он не достигает единицы и при полной связи признаков, а лишь стремится к единице при увеличении числа групп. Поэтому полезно провести корректировку коэффициента Пирсона, разделив его величину на *предельно возможное значение*, которое легко получается при подстановке в (11.5) значений $n_{ij} = n_{i.} = n_{.j}$, что имеет место при полной связи признаков.

Вычислим предельное значение $P_{\max}$ для разного числа категорий таблицы сопряженности m (табл. 11.8).

Таблица 11.8

Предельные значения коэффициента Пирсона

m	2	3	4	5	6	7	8	9	10
P	0,707	0,816	0,866	0,894	0,913	0,926	0,935	0,943	0,949

Так что коэффициент взаимной сопряженности Пирсона стремится к единице при увеличении числа групп. По данным табл. 11.7 при 5 группах корректированный показатель связи Пирсона составит:

$$P_{\text{корр.}} = 0{,}8253 : 0{,}894 = 0{,}923.$$

Величина коэффициента P зависит от числа категорий переменных, т.е. от числа строк и столбцов таблицы. Значение P не зависит от порядка следования строк и столбцов, а зависит только от значений самих клеточных частот.

Более совершенная мера связи предложена русским статистиком А. А. Чупровым. Рассматривая случай полной сопряженности, когда все нулевые частоты расположены только в диагональных клетках таблицы, т.е. каждому значению признака x соответствует определенное значение признака y, А. А. Чупров показал, что для квадратной таблицы $m \times m$ (т.е. число строк равно числу столбцов) показатель средней квадратической сопряженности при полной связи выражается формулой

$$\varphi^2 = m - 1. \qquad (11.6)$$

Если в формулу коэффициента взаимной сопряженности (11.4) подставим выражение φ^2, то получим формулу максимального значения P:

$$P_{\max} = \sqrt{\frac{m-1}{m}}. \qquad (11.7)$$

Из формулы (11.7) видно, что в случае полной связи предельное значение коэффициента взаимной сопряженности Пирсона в квадратных таблицах $m \times m$ зависит только от числа выделенных категорий по x и по y.

В общем случае, когда число строк не равно числу столбцов, $m \neq p$, Чупров предложил заменить выражение $\varphi^2 = m - 1$ средним геометрическим из числа степеней свободы для данной таблицы сопряженности, т.е.

$$\varphi^2 = \sqrt{d.f.} = \sqrt{(m-1)(p-1)}. \qquad (11.8)$$

Тогда мера тесноты связи Пирсона преобразуется в *коэффициент взаимной сопряженности Чупрова*:

$$T = \sqrt{\frac{\varphi^2}{\sqrt{(m-1)(p-1)}}} = \sqrt{\frac{\frac{\chi^2}{n}}{[(m-1)(p-1)]^{1/2}}}, \qquad (11.9)$$

$0 \leqslant T \leqslant 1$; $T = 0$ при независимости признаков, когда $\varphi^2 = 0$; $T = 1$ при полной связи признаков, когда $\varphi^2 = m - 1$. По сути, мера связи А. А. Чупрова основана на сопоставлении фактической средней квадратической сопряженности φ^2 с максимально возможной величиной $\varphi^2_{\max}$, т.е. характеризует долю фактической сопряженности признаков в их полной сопряженности. По данным табл. 11.7

$$T = \sqrt{\frac{2{,}1364}{[(5-1)(5-1)]^{1/2}}} = 0{,}731.$$

Квадрат коэффициента взаимной сопряженности А. А. Чупрова T^2 имеет смысл коэффициента детерминации. Для табл. 11.7 $T^2 = 0{,}534$, т.е. фактическая сопряженность составляет 53,4% полной сопряженности вероисповедания отца и матери. Величина $(1 - T^2)$ измеряет отклонение от полной сопряженности. В данном случае оно составляет почти 47%. В случае таблицы 2×2 выражение коэффициента взаимной сопряженности А. А. Чупрова совпадает с выражением коэффициента контингенции (11.1).

Коэффициент взаимной сопряженности Чупрова может достигать предельного значения, равного единице, только в случае квадратной таблицы ($m = p$): чем более несимметрична таблица, тем больше отличается T от единицы при полной связи признаков.

Для случая неквадратных таблиц, когда $m \neq p$, шведский математик и статистик Г. Крамер в 1946 г. предложил в формуле коэффициента взаимной сопряженности (11.11) учитывать минимальную из величин: либо число строк, либо число столбцов. Коэффициент взаимной сопряженности Г. Крамера имеет вид:

$$V = \sqrt{\frac{\varphi^2}{\min\{m-1, p-1\}}}. \qquad (11.10)$$

Очевидно, что в случае квадратной таблицы коэффициенты взаимной сопряженности А. А. Чупрова и Г. Крамера сов-

падают: если $m = p$, то $T = V$. Это замечание относится к рассмотренному примеру, где $m = p = 5$.

Как правило, показатель А. А. Чупрова гораздо строже оценивает тесноту связи, чем показатель К. Пирсона, слишком быстро приближающийся к единице.

Модифицируя для любого числа групп ранее предложенный для двух групп способ и формулы (11.2) и (11.3) с учетом обозначений частот f, получаем:

$$\eta^2 = \frac{\sum_{i=1}^{k} f_{ij}(i=j) - \sum_{i=1}^{k} f_{ij}'(i=j)}{\sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij} - \sum_{i=1}^{k} f_{ij}'(i=j)}, \qquad (11.11)$$

где

$$f_{ij}'(i=j) = \frac{f_i \cdot f_j (i=j)}{\sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij}},$$

т.е. частоты в клетках первой диагонали при отсутствии связи признаков. Подставив значения f_{ij}' в (11.11), получаем формулу, аналогичную (11.3):

$$\eta^2 = \frac{\left(\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f_{ij}\right)\sum_{i=1}^{k} f_{ij}(i=j) - \sum_{i=1}^{k} f_i \cdot f_j (i=j)}{\left(\sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij}\right)^2 - \sum_{i=1}^{k} f_i \cdot f_j (i=j)}, \qquad (11.12)$$

где

$$f_{ij}'(i=j) = \frac{\sum f_i \cdot \sum f_j (i=j)}{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f_{ij}},$$

т.е. $f_{ij}'(i=j)$ — это значения частот в клетках первой диагонали таблицы, которые были бы при отсутствии связи (пропорциональные частоты).

По данным табл. 11.7 имеем:

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f_{ij}'(i=j) = \frac{(214{,}1 \cdot 235{,}6) + (260{,}2 \cdot 273) +}{680{,}1}$$

$$\frac{+ (13{,}6 \cdot 13{,}5) + (67{,}8 \cdot 64{,}9) + (124{,}4 \cdot 93{,}1)}{680{,}1} = 202{,}17.$$

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f_{ij}(i=j) = 146{,}1 + 195{,}9 + 10{,}5 + 62{,}8 + 77{,}7 = 493{,}0.$$

$$\eta^2 = \frac{493 - 202{,}17}{680{,}1 - 202{,}17} = 0{,}6085; \qquad \eta = 0{,}780.$$

Таким образом, за счет предпочтения браков между лицами одного вероисповедания на главную диагональ «собралось» 60,85% возможных родительских пар сверх равномерного распределения: связь составила 60,85% от предельно тесной. Итак, все способы измерения показали, что влияние религии на формирование супружеских пар в ФРГ в 1993 г. было значительное.

Очевидно, что коэффициенты взаимной сопряженности — симметричные меры связей, т.е. $T_{yx} = T_{xy}$. Все они используются для измерения тесноты связи после того, как факт наличия связи доказан на основе критерия хи-квадрат.

11.4. Теоретико-информационные меры связей

Коэффициенты взаимной сопряженности основаны на критерии хи-квадрат. Следовательно, их можно использовать, когда выполняются все предпосылки применения хи-квадрата: большой объем наблюдений, большое число строк и столбцов таблицы сопряженности, теоретически частоты составляют не менее 5 единиц, $\hat{n}_{ij} \geqslant 5$. При выполнении этих условий даже нулевое значение коэффициента взаимной сопряженности может не означать независимости признаков.

Отмеченные ограничения отсутствуют у теоретико-информационных мер связей, основанных на величине количества информации $I(y, x)$. С этой целью оценивается неопределенность распределения переменной y (без учета знания переменной x), т.е. вычисляется *полная энтропия распределения переменной* y:

$$H(y) = -\sum_{(j)} p(y_j)\log_2 p(y_j), \qquad (11.13)$$

где j — номер категории переменной y;

$p(y_j)$ — вероятность (частость) появления j-го значения переменной y.

Существуют таблицы величин — $p(y_j)\log_2 p(y_j)$ (табл. П.13 приложения), которые значительно упрощают расчеты. Полная энтропия распределения вычисляется на основе *безусловного* распределения. После чего рассчитывается неопределенность распределения y при закрепленном значении x, т.е. энтропия *условного* распределения y:

$$H_{x_i}(y) = -\sum_{(i)} p_{x_i}(y_j)\log_2 p_{x_i}(y_j), \qquad (11.14)$$

где $p_{x_i}(y_j) = \frac{n_{ij}}{n_{i.}}$.

Формула (11.14) определяет энтропию распределения y при i-м значении переменной x. В целом условная энтропия определяется как:

$$H_x(y) = \sum_{(i)} H_{x_i}(y)p(x_i). \qquad (11.15)$$

Следовательно, энтропия распределения переменной y с учетом знания переменной x определяется как средняя взвешенная из энтропий условных распределений переменной y при i-м значении переменной x. Если x полностью предопределяет распределение y, то $H_x(y) = 0$, т.е. знание переменной x полностью устраняет неопределенность наших знаний об y. Если x не связан с y, то $H_x(y) = H(y)$.

Разность между полной и условной энтропией переменной y есть не что иное, как *количество информации* о переменной y за счет знания переменной x, $I(y, x)$. Количество информации, так же как и энтропия распределения, измеряется в *битах*. Это информационная мера, основанная на двоичной системе записи информации: 0 и 1. Энтропия распределения изменяется от 0 до H_{max}. Нулевое значение энтропии распределения означает, что неопределенности нет: все единицы принадлежат к одной и той же категории переменной y. Максимальная неопределенность соответствует равновероятному

распределению. Частости такого распределения $p(y_i) = \frac{1}{k}$, где k — число категории переменной y ($k = 1, ..., p$).

Разработано целое семейство теоретико-информационных коэффициентов связи. Наиболее популярен коэффициент нормированной информации:

$$R_{y/x} = \frac{H(y) - H_x(y)}{H(y)} = \frac{I(y, x)}{H(y)}. \qquad (11.16)$$

Очевиден смысл этого показателя как меры относительной редукции неопределенности наших знаний об y при получении знания об x. Коэффициент нормированной информации $R_{y/x}$ обладает следующими свойствами: 1) $0 \leqslant R_{y/x} \leqslant 1$; 2) $R_{y/x} = 0$, если переменные независимы; 3) $R_{y/x} = 1$, если между y и x имеет место полная (функциональная) связь; 4) $R_{y/x}$ инвариантен к перестановке местами строк и столбцов таблицы сопряженности; 5) $R_{y/x}$ инвариантен по отношению к значениям переменных, он определяется только на основе вероятностей (частостей) распределения. По сути, этот коэффициент аналогичен коэффициенту детерминации: в числителе — объясненная дисперсия (здесь — информация); в знаменателе — полная дисперсия (здесь — энтропия безусловного распределения). Очевидно, что $0 \leqslant R_{y/x} \leqslant 1$.

Рассчитывая этот коэффициент, можно видеть, что он очень медленно возрастает при повышении тесноты связи: для случаев, когда коэффициенты взаимной сопряженности оказываются равными 0,3—0,4, теоретико-информационный коэффициент связи, рассчитанный по тем же данным, составит примерно 0,10—0,12. Это нужно принимать во внимание, если используются разные меры связи и делается оценка тесноты связи для принятия управленческих решений.

Пример. Изучается зависимость образования взрослых сына или дочери от образования матери. По данным табл. 11.9 вычислим полную энтропию переменной y по безусловному распределению, т.е. по данным о распределении сына или дочери по уровню образования без учета образования матери:

$$H_{(y)} = (0{,}384 \cdot \log_2 0{,}384) + (-0{,}311 \cdot \log_2 0{,}311) + (0{,}305 \cdot \log_2 0{,}305) = 1{,}5767 \text{ бита}.$$

Таблица 11.9

Зависимость образования молодого поколения от образования матери

Образование матери x	Образование сына или дочери, y			Всего	
	Высшее	Среднее специальное	Общее среднее и неполное среднее	итого	в процентах к итогу
Высшее	83,5	10,4	6,1	100	12,4
Среднее специальное	42,5	50,5	7,0	100	14,8
Общее среднее	55,0	25,4	19,6	100	16,4
Неполное среднее	26,2	36,9	36,9	100	20,6
Начальное	20,2	29,6	50,2	100	35,8
Итого	38,4	31,1	30,5	100	100

По характеру итогового распределения видим, что оно близко к равновероятному, следовательно, полученное значение энтропии близко к H_{max}. Затем рассчитаем энтропии условных распределений детей по образованию при условии определенного образования матери:

$$H_{x_1}(y) = (-0{,}835 \cdot \log_2 0{,}835) + (-0{,}104 \cdot \log_2 0{,}104) + (-0{,}061 \cdot \log_2 0{,}061) = 0{,}8031 \text{ бита.}$$

Точно так же рассчитываются значения энтропии других условных распределений детей по образованию:

$$H_{x_2}(y) = 1{,}3049 \text{ бита; } H_{x_3}(y) = 1{,}4374 \text{ бита,}$$

$$H_{x_4}(y) = 1{,}5677 \text{ бита, } H_{x_5}(y) = 1{,}4851 \text{ бита.}$$

Сравнение полученных значений $H_{x_i}(y)$ свидетельствует о влиянии образования матери. Так, в первом случае при наличии высшего образования у матери энтропия распределения детей наименьшая, т.е. если мать имеет высшее образование, то более вероятно, что сын или дочь тоже получит высшее образование.

Условная энтропия распределения детей по образованию вычисляется как средняя взвешенная из полученных значений $H_{x_i}(y)$:

$$H_x(y) = 0{,}8031 \cdot 0{,}124 + 1{,}3049 \cdot 0{,}148 + 1{,}4374 \cdot 0{,}164 + 1{,}5677 \cdot 0{,}206 + 1{,}4851 \cdot 0{,}358 = 1{,}3830 \text{ бита.}$$

Данные об образовании матери уменьшили неопределенность знания об образовании детей. Полученное количество информации равно:

$$I(y, x) = 1{,}5767 - 1{,}3830 = 0{,}1937 \text{ бита.}$$

Коэффициент нормированной информации равен:

$$R_{y/x} = 0{,}1937/1{,}5767 = 0{,}123.$$

Если бы такое значение принял коэффициент корреляции, мы бы сделали вывод, что связь слабая. А применяя коэффициент нормированной информации, следует помнить, что, если $R_{y/x} \geqslant 0{,}1$, связь либо умеренно тесная, либо тесная.

Как уже отмечалось, по своему строению коэффициент нормированной информации аналогичен коэффициенту детерминации, так что его можно выражать в процентах. В данном примере количество информации составляет 12,3% энтропии распределения детей по уровню образования.

Коэффициент нормированной информации является асимметричной мерой связи: $R_{y/x} \neq R_{x/y}$.

Симметризованный коэффициент информации имеет вид:

$$R(y, x) = \frac{H(y) - H_x(y)}{1/2[H(y) + H(x)]}, \qquad (11.17)$$

где $H(x)$ — энтропия распределения по переменной x (для безусловного распределения).

Этот коэффициент имеет те же свойства, что и $R_{y/x}$.

По данным табл. 11.9 получаем

$$R(y, x) = \frac{0{,}1937}{\frac{1}{2}(2{,}209 + 1{,}5767)} = 0{,}104.$$

Теоретико-информационные меры связей используются, когда в клетках таблицы мало единиц ($n_{ij} < 5$), когда имеются пустые клетки. Эти меры связи не предъявляют никаких требований к исходным данным.

11.5. Другие меры связей между номинальными переменными

Рассмотрим меры связей λ-Гутмана и τ-Гудмена и Краскала. Это семейства мер связи, включающие асимметричные меры и симметричную меру связи: λ_b, λ_a, λ и τ_b, τ_a, τ.

Меры связи λ-Гутмана имеют очень простую структуру. Они определяются по таблицам, где хотя бы одна переменная номинальная и переменные недихотомические. Поскольку эти меры разработаны для номинальных переменных, знаки при λ не имеют значения (плюс или минус).

Асимметричные меры λ_b и λ_a зависят от того, какая из переменных рассматривается как зависимая.

Статистика λ_b основана на сравнении двух ситуаций: определение категории *B*, когда: 1) нет никакой дополнительной информации, 2) когда известна *A*-категория наблюдения. Мера λ_b характеризует относительный прирост вероятности предсказания *B*-категории при переходе от первой ситуации ко второй:

$$\lambda_b = \frac{\sum_{(j)} n_{ij(\max)} - n_{j(\max)}}{n - n_{j(\max)}}, \qquad (11.18)$$

где $n_{ij(\max)}$ — максимальная клеточная частота в *i*-й строке;
n_j — максимальная маргинальная частота (максимальный итог по столбцу);
n — объем выборки.

Пример. Рассмотрим порядок расчета этих мер.

По данным табл. 11.10 получаем:

$$\lambda_b = \frac{(506 + 58 + 48) - 564}{1146 - 564} = \frac{48}{582} = 0{,}082,$$

т.е. знание намерений избирателей на 8,2% повышает наше знание о знакомстве с программами кандидатов в депутаты.

Аналогично определяется статистика λ_a:

$$\lambda_a = \frac{\sum_{(j)} n_{ij(\max)} - n_{i(\max)}}{n - n_{i(\max)}}. \qquad (11.19)$$

Таблица 11.10

Распределение опрошенных по намерениям участвовать в выборах депутатов ЗАКС

Переменная *A*	Переменная *B*			Итого
	знаком с программами кандидатов	не знаком с программами кандидатов	затрудняюсь ответить	
Намерен участвовать в выборах	506	375	18	899
Не намерен участвовать в выборах	25	58	53	136
Затрудняюсь ответить	33	30	48	111
Итого	564	463	119	1146

В числителе λ_a показана сумма максимальных клеточных частот по столбцам; $n_{i(\max)}$ — максимальный итог по строке.

По данным примера получаем:

$$\lambda_a = \frac{(506 + 375 + 53) - 899}{1146 - 899} = \frac{35}{247} = 0{,}142,$$

т.е. знакомство с программами кандидатов на 14,2% повышает знание о намерении участвовать в выборах.

Статистика λ определяется как объединение λ_a и λ_b:

$$\lambda = \frac{\left[\sum_{(j)} n_{ij(\max)} - n_{j(\max)}\right] + \left[\sum_{(i)} n_{ij(\max)} - n_{i(\max)}\right]}{2n - n_{j(\max)} - n_{i(\max)}}. \qquad (11.20)$$

По данным примера

$$\lambda = \frac{48 + 35}{582 + 247} = \frac{83}{829} = 0{,}100,$$

т.е. знание обеих переменных на 10% повышает вероятность предсказания принадлежности респондента к той или иной категории. Статистика λ используется в том случае, когда трудно поставить переменные в отношение «зависимая — независимая». Меры λ-Гутмана имеют тот недостаток, что они принимают нулевое значение, если все максимальные кле-

точные частоты оказываются в одном и том же столбце или в одной и той же строке таблицы. При этом $\sum_j n_{ij(\max)} = n_{j(\max)}$ или $\sum_i n_{ij(\max)} = n_{i(\max)}$. Тогда числители λ_b и λ_a равны 0, но это может не означать независимости переменных.

Для того чтобы избежать таких случаев, Л. Гудмен и Э. Краскал предложили меры τ (τ_b, τ_a, τ).

Все τ-статистики основаны на предсказании различных категорий переменных A и B пропорционально наблюдаемым итогам распределения этих переменных:

$$\tau_b = \frac{\sum_{(i)}\sum_{(j)}[(nn_{ij} - n_{j.}n_{i.})^2 / n_{i.}]}{n\left(n^2 - \sum_{(j)} n_{.j}^2\right)}. \qquad (11.21)$$

По данным табл. 11.10 получим:

$$\tau_b = \frac{1}{1146(1146^2 - 564^2 - 463^2 - 119^2)} \cdot \left[\frac{(1146 \cdot 506 - 564 \cdot 899)^2}{899} + \right.$$

$$+ \frac{(1146 \cdot 375 - 463 \cdot 899)^2}{899} + \ldots + \frac{(1146 \cdot 30 - 463 \cdot 111)^2}{111} +$$

$$\left. + \frac{(1146 \cdot 48 - 119 \cdot 111)^2}{111}\right] = 0{,}08.$$

Знание того, известны ли избирателям программы кандидатов, на 8% повышает вероятность предсказания участия в выборах респондентов.

Аналогично определяются τ_a; τ, как и λ, определяется объединением τ_b и τ_a.

Меры связи τ ближе по своей конструкции к статистике хи-квадрат, являются более надежными мерами, нежели λ.

За последние 100 лет учеными разных стран разработано множество мер измерения связей категоризованных переменных, с которыми можно познакомиться, изучив литературу, указанную в конце данной главы.

11.6. Коэффициенты корреляции рангов

К мерам тесноты парной связи относится и предложенный английским психологом Ч. Спирменом (1863—1945) *коэффициент корреляции рангов. Ранги* — это порядковые номера единиц совокупности в ранжированном ряду. Если проранжировать совокупность по двум признакам, связь между которыми изучается, то полное совпадение рангов означает максимально тесную прямую связь, а полная противоположность рангов — максимально тесную обратную связь. Ранжировать оба признака необходимо в одном и том же порядке: либо от меньших значений признака к большим, либо наоборот. Если ранг единиц совокупности по признакам x и y обозначить как p_{x_i}, p_{y_i}, то коэффициент корреляции рангов имеет вид:

$$r_{p_x p_y} = \frac{\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)(p_{y_i} - \bar{p}_y)}{\sqrt{\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)^2 \sum_{i=1}^{n}(p_{y_i} - \bar{p}_y)^2}}, \qquad (11.22)$$

где $\bar{p}_x = \bar{p}_y$ — средние ранги в ряду натуральных чисел от 1 до n, равные, как известно, $(n + 1)/2$.

Известно, что сумма квадратов отклонений чисел натурального ряда от их средней величины $\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)^2$ и $\sum_{i=1}^{n}(p_{y_i} - \bar{p}_y)^2$ равна $(n^3 - n)/12$. Следовательно, знаменатель формулы (11.24) есть $(n^3 - n)/12$.

Рассмотрим далее разности рангов $d_i = p_{x_i} - p_{y_i}$ и сумму квадратов разностей:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(p_{x_i} - p_{y_i})^2 = \sum_{i=1}^{n}[(p_{x_i} - \bar{p}_x) - (p_{y_i} - \bar{p}_y)]^2 =$$

$$= \sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)^2 + \sum_{i=1}^{n}(p_{y_i} - \bar{p}_y)^2 - 2\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)(p_{y_i} - \bar{p}_y) =$$

$$= \frac{2(n^3 - n)}{12} - 2\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)(p_{y_i} - \bar{p}_y).$$

Отсюда:

$$\sum_{i=1}^{n}(p_{x_i}-\bar{p}_x)(p_{y_i}-\bar{p}_y)=\left(\frac{2(n^3-n)}{12}-\sum_{i=1}^{n}d_i^2\right):2=\frac{n^3-n}{12}-\frac{\sum_{i=1}^{n}d_i^2}{2}.$$

Это числитель коэффициента корреляции рангов. Подставив в (11.24) найденные выражения для числителя и для знаменателя, имеем:

$$r_{p_xp_y}=\frac{\frac{(n^3-n)}{12}-\frac{\sum_{i=1}^{n}d_i^2}{2}}{\frac{n^3-n}{12}}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n^3-n}. \qquad (11.23)$$

Это и есть *формула Спирмена*. Обычно коэффициент корреляции рангов Спирмена (или коэффициент ранговой корреляции) обозначается буквой ρ («ро», греч).

Примущество коэффициента корреляции рангов состоит в том, что ранжировать можно и по таким признакам, которые нельзя выразить численно: можно проранжировать кандидатов на занятие определенной должности по профессиональному уровню, по умению руководить коллективом, по личному обаянию и т.п. При экспертных оценках можно ранжировать оценки разных экспертов и найти их корреляции друг с другом, чтобы затем исключить из рассмотрения оценки эксперта, слабо коррелированные с оценками других экспертов. Коэффициент корреляции рангов применяется для оценки устойчивости тенденции динамики (см. подразд. 12.9).

Недостатком коэффициента корреляции рангов является то, что одинаковым разностям рангов могут соответствовать совершенно отличные разности значений признаков (в случае количественных признаков). Поэтому для последних следует считать корреляцию рангов, как и коэффициент знаков Фехнера, приближенными мерами тесноты связи, обладающими меньшей информативностью, чем коэффициент корреляции числовых значений признаков.

Рассчитаем коэффициент корреляции рангов по данным табл. 11.11. Ранги присвоены в соответствии со значениями переменных (см. табл. 9.1).

Таблица 11.11

Расчет коэффициента корреляции рангов (по данным табл. 9.1)

Номер хозяйства	Ранг по затратам на одну голову p_x	Ранг по надою молока p_y	$d=p_x-p_y$	d^2
1	7	10	−3	9
2	1	1	0	0
3	2	3	−1	1
4	13	6	7	49
5	6	9	−3	9
6	3	5	−2	4
7	4	4	0	0
8	5	7	−2	4
9	9	2	7	49
10	14	14	0	0
11	11	12	−1	1
12	12	11	1	1
13	10	13	−3	9
14	8	8	0	0
15	16	15	1	1
16	15	16	−1	1
Σ	136	136	0	138

Коэффициент корреляции рангов по формуле Спирмена

$$r_{p_xp_y}=1-\frac{6\cdot 138}{16^3-16}=0{,}797.$$

Полученное значение больше коэффициента Фехнера (0,75), но намного меньше обычного коэффициента корреляции, составившего 0,916. Как видим, недоучет размеров отклонений признаков от их средних величин занижает меру тесноты связи.

Если среди значений рангов признаков x и y встречается несколько одинаковых, образуются *связанные ранги*, т.е. одинаковые средние номера; например, вместо одинаковых по порядку третьего и четвертого значений признака будут два ранга по 3,5. В таком случае коэффициент Спирмена вычисляется как:

$$r_{p_xp_y}=1-\frac{6\sum d^2-A-B}{\sqrt{(n^3-n-12A)(n^3-n-12B)}}, \qquad (11.24)$$

где

$$A = \frac{1}{12}\sum_j (A_j^3 - A_j); \qquad B = \frac{1}{12}\sum_k (B_k^3 - B_k);$$

j — номера связок по порядку для признака x;
A_j — число одинаковых рангов в j-й связке по x;
k — номера связок по порядку для признака y;
B_k — число одинаковых рангов в k-й связке по y.

Значимость коэффициента корреляции рангов Спирмена проверяется на основе t-критерия Стьюдента:

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}}.$$

Если $t > t_{\alpha,\ d.f.\ =\ n-2}$, значение ρ статистически значимо.

Коэффициент корреляции рангов может быть рассчитан и по формуле, предложенной английским статистиком М. Кендэлом:

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}, \qquad (11.25)$$

где S — фактическая сумма рангов;
$\frac{1}{2}n(n-1)$ — максимальная сумма рангов.

Этот коэффициент так же, как и ρ, изменяется в пределах $-1 < \tau < 1$. Он дает несколько более строгую оценку связи, нежели коэффициент Спирмена: $\rho_s \approx \frac{3}{2}\tau$. Это соотношение выполняется при большом числе наблюдений, $n > 30$, и слабых либо умеренно тесных связях. Для расчета τ все единицы ранжируются по признаку x; по ряду другого признака y подсчитываются для каждого ранга число последующих рангов, превышающих данный (их сумму обозначим P), и число последующих рангов ниже данного (их сумму обозначим Q). Тогда: $S = P - Q$. Можно показать, что $P + Q = \frac{1}{2}n(n-1)$, так что τ может быть представлен как

$$\tau = \frac{P-Q}{P+Q}. \qquad (11.26)$$

Вычислим коэффициент корреляции рангов Кендэла по данным табл. 11.12.

Таблица 11.12

Ранжирование данных по переменным x и y

Ранги по x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ранги по y	1	3	5	4	7	9	10	8	2	13	12	11	6	14	16	15

$$S = (15-0) + (13-1) + (11-2) + (11-1) + (9-2) + (7-3) + (6-3) + (6-2) + (7-0) + (3-3) + (3-2) + (3-1) + (3-0) + (2-0) + (0-1) = 99 - 21 = 78;$$

$$\tau = \frac{78}{\frac{1}{2}16 \cdot 15} = 0{,}65.$$

Хотя $\tau < r_s$ ($r_s = 0{,}797$), но поскольку связь тесная и $n < 30$, соотношение между этими двумя коэффициентами не вполне соответствует упомянутому: коэффициент Спирмена в нашем примере превосходит τ не в 1,5 раза, а на 23%.

11.7. Коэффициент конкордации

Коэффициент конкордации — характеристика связи между несколькими признаками, измеренными по порядковой шкале.

Этот коэффициент вычисляется по формуле

$$W = \frac{S}{\frac{1}{12}[m^2(n^3-n)]}, \qquad (11.27)$$

где S — сумма квадратов отклонений суммы рангов каждого объекта от средней суммы рангов;
m — количество порядковых переменных;
n — объем выборки.

Коэффициент принимает значения от 0 до 1, $0 \leqslant W \leqslant 1$. По своей сути коэффициент конкордации — среднее значение из коэффициентов ранговой корреляции Спирмена между каждой парой рангов. Число таких коэффициентов равно числу сочетаний из m по 2, т.е. C_m^2.

Значениям каждой переменной приписываются ранги. Ранг 1 устанавливается наименее важному значению: минимальному — для стимулянт, т.е. для переменных типа «чем больше, тем лучше», и максимальному для дестимулянт, т.е.

для переменных типа «чем больше, тем хуже». Если нельзя отдать предпочтение нескольким объектам, то каждому из них присваивается средний ранг, определяемый как средний арифметический из суммы соответствующих мест («связанные ранги»). Скажем, если нельзя отдать предпочтение второму, третьему и четвертому объектам, то каждому из этих объектов присваивается ранг, равный $\frac{2+3+4}{3} = 3$. Связанные ранги могут быть и дробными.

Значимость W_a проверяется на основе критерия χ^2:

$$\chi^2 = \frac{12S}{m \cdot n(n-1)}.$$

Полученное значение χ^2 сравнивается с критическим $\chi^2_{\alpha,\, d.f.\, =\, n-1}$.

В случае связанных рангов формула коэффициента конкордации имеет вид:

$$W = \frac{S}{\frac{1}{12}\left[m^2(n^3-n) - m\sum_{j=1}^{m} T_j\right]}, \qquad (11.28)$$

где T_j — характеристика связанности рангов по j-й переменной,

$$T_j = \frac{1}{12}\sum_{j=1}^{m}(t_j^3 - t_j);$$

t_j — количество связанных рангов по j-й переменной.

Знаменатель коэффициента конкордации представляет гипотетическую сумму рангов, получаемую в случае их полной согласованности. Чаще всего коэффициент конкордации используется для измерения согласованности мнений экспертов о влиянии различных признаков на результативную переменную.

Пример. Три эксперта дали характеристику разведанного месторождения газа по пяти признакам (табл. 11.13).

Числитель W: $S = 40{,}5$; знаменатель W: $= \frac{1}{12}[5^2(3^3 - 3) - 5 \cdot (\frac{1}{12}(2^3 - 2) \cdot 2)] = 49{,}583$. Величина коэффициента конкордации составила $W = \frac{40{,}5}{49{,}583} = 0{,}817$, т.е. согласованность мнений экспертов довольно высокая.

Таблица 11.13

Расчет коэффициента конкордации

Признак	Эксперты			Сумма рангов по каждому признаку	Квадрат отклонения от средней суммы рангов
	1-й	2-й	3-й		
Мощность месторождения	5	3	5	13	16
Трудности разработки	1	5	4	10	1
Наличие трудовых ресурсов	3,5	4	3	10,5	2,25
Потребность в оборудовании	3,5	1	1,5	6	9
Развитие транспортных коммуникаций	2	2	1,5	5,5	12,25
Итого	15	15	15	45	40,5
В среднем	3	3	3	9	—

Можно оценить значимость коэффициента конкордации по критерию хи-квадрат:

$$\chi^2 = \frac{12S}{mn(n-1) - \frac{1}{n-1}\sum_{j=1}^{m} T_j}, \qquad (11.29)$$

где $d.f. = (n-1)$.

Для нашего примера

$$\chi^2 = 16{,}47.$$

Поскольку $\chi^2_{\text{табл}}$ $(\alpha = 0{,}05,\ d.f. = 2) = 5{,}99$, *то можно сделать* вывод, что полученное значение коэффициента конкордации статистически значимо.

РЕЗЮМЕ

Способы измерения связей между признаками зависят от того, по какой шкале они измерены: номинальной, порядковой, интервальной или шкале отношений.

В собираемых статистических данных непрерывно возрастает доля нечисловой информации. Это объясняется несколькими причинами:

- стремлением учесть человеческий фактор (в бизнесе, потреблении), выявить ориентации и предпочтения людей;
- сбором информации в форме нечисловых данных с тем, чтобы не затронуть количественные показатели, составляющие коммерческую тайну;
- использованием рейтингов (банков, предприятий, учебных заведений, политических деятелей и т.д.).

Измерение связи между неколичественными переменными основано на таблице сопряженности — двух- или трехмерном распределении единиц совокупности. Если переменные дихотомические, то данные представляются в таблице 2 × 2 и вычисляются специальные меры связи: коэффициенты ассоциации, коэффициенты контингенции.

По таблицам сопряженности $m \times p$ вычисляются коэффициенты взаимной сопряженности, основанные на тестовой статистике хи-квадрата.

В случае, если нельзя выполнить условия применения статистики хи-квадрат, рекомендуется пользоваться теоретико-информационными мерами связей, основанными на измерении энтропии распределений и количества информации. В качестве мер связей между номинальными переменными используются меры связи: λ-Гутмана, τ-Гудмена и Краскала и др.

Корреляция между порядковыми переменными измеряется коэффициентом ранговой корреляции. Широко распространены коэффициенты ранговой корреляции Спирмена и Кендэла.

Меры связей между неколичественными переменными применяются при обработке данных экспертных опросов. Если экспертам нужно оценить объект не по одному, а по нескольким свойствам, то используется коэффициент конкордации.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Аптон Г.* Анализ таблиц сопряженности: Пер. с англ. — М.: Финансы и статистика, 1982.
2. *Елисеева И. И., Рукавишников В. О.* Логика прикладного статистического анализа. — М.: Финансы и статистика, 1982.
3. *Информатика* в статистике: Словарь-справочник. — М.: Финансы и статистика, 1994.
4. *Ниворожкина Л. И., Морозова З. А.* Сборник задач по математической статистике с элементами теории вероятностей. — Ростов-на-Дону: РИНХ, 2002.

Глава 12

СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ ДИНАМИКИ

12.1. Виды динамических рядов. Сопоставимость данных в изучении динамики

Одно из основных положений научной методологии — необходимость изучать все явления в развитии, во времени. Это относится и к статистике: она должна дать характеристику изменений статистических показателей во времени. Как изменяются год за годом валовой национальный продукт и национальный доход страны? Как возрастает или снижается уровень оплаты труда? Велики ли колебания урожайности зерновых культур и существует ли тенденция ее роста? Ответ на аналогичные вопросы может дать только специальная система статистических методов, предназначенная для изучения развития изменений во времени, или, как принято в статистике говорить, изучения динамики.

Изучение динамики того или иного объекта, явления начинается с построения ***ряда динамики***, или временного ряда (англ. time series). Динамический ряд — это таблица, в которой представлены значения показателя за последовательные периоды или на моменты времени. Каждое значение показателя называется *уровнем ряда*. Динамический ряд является *интервальным*, если каждый уровень представляет собой итог процесса за некоторый интервал времени (например, ряды в табл. 10.1; 10.4; 10.10, или *моментным*, если уровни отражают состояние объекта в отдельные моменты времени.

Важнейшим условием построения динамического ряда является *сопоставимость его уровней*. Бессмысленно изучать динамику выпуска продукции предприятием или в регионе,

если стоимость продукции разных лет выражена в различных ценах, растущих в результате инфляции. Объем продукции должен быть пересчитан в условно-постоянные цены.

Пример. Рассмотрим динамику валового регионального продукта (ВРП) Санкт-Петербурга:

	1996	1997		1998	1999	2000
ВРП Санкт-Петербурга, млрд руб.	66 332	75 735	млн руб.	92 029	150 727	205 400

Данные ВРП по годам несопоставимы из-за инфляции и финансового кризиса 1998 г. Это отражается и в различии единиц счета (миллиарды рублей – до 1998 г., миллионы рублей — после 1998 г.).

Представление о динамике ВРП дают темпы его изменения в сопоставимых ценах (в процентах к предыдущему году):

	1997	1998	1999	2000
Изменение ВРП в сопоставимых ценах, в процентах к предыдущему году	98,6	94,7	106,0	110,5

Уровни валового сбора зерна в области (т.е. собранного урожая) должны быть сопоставимы по территории: если границы области на протяжении изучаемого периода изменялись, то динамика уровней не отразит развитие производства зерна. Необходимо показать динамику валового сбора на одной и той же территории. Все уровни должны быть выражены в одинаковых единицах измерения. Они должны быть учтены или рассчитаны по единой методике. При изменении методики производится пересчет уровней предыдущих периодов по новой методике расчета. Например, с 1999 г. Госкомстат России перешел на единую с Европейским союзом и ООН методику определения урожайности сельскохозяйственных культур, которая заключается в делении валового показателя сбора на фактически убранную площадь. Ввиду этого ранее рассчитанные показатели урожайности на весеннюю продуктивную площадь подлежат пересчету.

Проблема сопоставимости уровней динамического ряда весьма сложна, особенно при изучении выпуска промышленной продукции, ассортимент которой часто изменяется. Бессмысленно, например, измерять темп развития производства

телевизоров или персональных компьютеров по данным их выпуска в тысячах штук, ведь главное в развитии высокотехнологичных отраслей – совершенствование качества продукции. В значительной степени то же относится к производству станков, автомобилей, самолетов. Не следует абсолютизировать и требование территориальной сопоставимости уровней. Например, если изучается динамика населения города, то было бы неверно брать данные по постоянной территории. Расширение территории города является необходимой составляющей его развития, и нужно показывать в разные годы то население, которое проживало в фактических (меняющихся) границах. Таким образом, кроме общих положений о сопоставимости уровней динамического ряда, в каждом конкретном исследовании необходимо добиваться соблюдения конкретных условий сопоставимости.

12.2. Элементы динамики: основная тенденция и колебания

Рассмотрим данные табл. 12.1. Условимся, что относящиеся к отдельным годам значения урожайности картофеля будем называть *уровнями*, а всю их последовательность с 1989 по 1999 г. — рядом динамики.

Ряд динамики состоит из двух строк или столбцов: промежутков или моментов времени, к которым относятся уровни, и самих уровней признака (показателя). Ряд, в котором время задано в виде промежутков — лет, месяцев, суток, называется *интервальным динамическим рядом*. Ряд, в котором время задано в виде конкретных дат (моментов времени), называется *моментным динамическим рядом*. Например, ряд численности населения по оценке на 1 января каждого года.

Таблица 12.1

Динамика урожайности картофеля в хозяйстве

Показатель	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Урожайность, ц/га	149	145	168	146	177	176	190	186	176	211	170

Вернемся к табл. 12.1. Сравнивая уровни разных лет, мы замечаем, что в целом урожайность возрастает. Однако нередко уровень урожайности следующего года оказывается ниже уровня предыдущего. Иногда рост по сравнению с предыдущим годом велик, как в 1990 г., а иногда мал. Следовательно, рост урожайности наблюдается лишь в среднем, как *тенденция*. В отдельные же годы уровни испытывают *колебания*, отклоняясь от основной тенденции. Эти колебания урожайности связаны в основном с различием метеорологических условий в разные годы.

Если рассматривать динамические ряды месячных уровней производства мяса и молока, ряды объема продажи разных видов одежды и обуви, ряды заболеваемости населения, выявятся регулярно повторяющиеся из года в год *сезонные* колебания уровней. В силу солнечно-земных связей частота полярных сияний, интенсивность гроз, те же изменения урожайности отдельных сельскохозяйственных культур и ряд других процессов имеют циклическую 10—11-летнюю колеблемость. Колебания числа рождений, связанные с потерями в войне, повторяются с угасающей амплитудой через поколение, т.е. через 20—25 лет.

Тенденция динамики связана с действием *долговременно* существующих причин и условий развития, хотя, конечно, после какого-то периода эти причины и условия тоже могут измениться и породить уже другую тенденцию развития изучаемого объекта. Колебания же, напротив, связаны с действием *краткосрочных,* или циклических, факторов, влияющих на отдельные уровни динамического ряда и отклоняющих уровни от тенденции то в одном, то в другом направлении. Например, тенденция динамики урожайности связана с прогрессом агротехники, с укреплением экономики данной совокупности хозяйств, совершенствованием организации производства. Колеблемость урожайности вызвана чередованием благоприятных и неблагоприятных по погоде лет, циклами солнечной активности, колебаниями в развитии вредных насекомых и болезней растений.

При статистическом изучении динамики необходимо четко разделить два ее основных элемента — тенденцию и колеблемость, чтобы дать каждому из них количественную характеристику с помощью специальных показателей. Смешение



Рис. 12.1. Динамика урожайности картофеля

тенденции и колеблемости ведет к неверным выводам. Если из табл. 12.1 произвольно взять данные за отдельные годы и сравнить их друг с другом, можно получить выводы, прямо противоположные истине. Например, если сравнить урожайность в 1998 г. с урожайностью в 1990 г., то получим, что за 8 лет она возросла на 66 ц/га, т.е. более чем по 8 ц/га за год. Если же урожайность в 1999 г. сравнить с ее уровнем в 1991 г., то получим, что за 8 лет, из которых 7 лет те же, что и в предыдущем сравнении, урожайность возросла всего лишь на 2 ц/га.

Тенденцию и колебания наглядно показывает график (рис. 12.1). По оси абсцисс всегда отражается время, по оси ординат — уровни. По обеим осям строго соблюдается масштаб, иначе характер динамики будет искажен.

На рис. 12.1 хорошо заметно, что рост урожайности в 1989—1999 гг. характеризовался линейной тенденцией, а колеблемость была хаотической, без явной цикличности. О линии тренда и ее уравнении будет сказано ниже.

12.3. Показатели, характеризующие тенденцию динамики

Для того чтобы построить систему показателей, характеризующих тенденцию динамики, нужно ответить на вопрос: какие черты, свойства этой тенденции необходимо измерить

и выразить в статистических показателях? Очевидно, нас интересует величина изменений уровня как в абсолютном, так и в относительном выражении (на какую долю, процент уровня, принятого за базу, произошло изменение?). Далее нас интересует: является ли изменение равномерным или неравномерным, ускоренным или замедленным? Наконец, нас интересует выражение тенденции в форме некоторого достаточно простого уравнения, наилучшим образом аппроксимирующего фактическую тенденцию динамики. Понятие об уравнении тенденции динамики было введено в статистику английским ученым Гукером в 1902 г. Он предложил называть такое уравнение *трендом* (англ. the trend — направление, тенденция).

Для того чтобы нагляднее представить показатели, характеризующие тенденцию, следует абстрагироваться от колеблемости и выявить динамический ряд в форме «чистого» тренда при отсутствии колебаний. Пример такого ряда представлен в табл. 12.2.

Абсолютное изменение уровней — в данном случае его можно назвать *абсолютным приростом* — это разность между сравниваемым уровнем и уровнем более раннего периода, принятым за базу сравнения. Если эта база — непосредственно предыдущий уровень, показатель называют *цепным*, если за базу взят, например, начальный уровень, показатель называют *базисным*. Формулы абсолютного изменения уровня:

$$\text{цепное: } \Delta_i = y_i - y_{i-1};$$
$$\text{базисное: } \Delta_i = y_i - y_0. \qquad (12.1)$$

Таблица 12.2

Абсолютные и относительные показатели тенденции

Номер периода или момента времени	Уровни ряда, тыс. т	Абсолютное изменение уровней, тыс. т в год	Ускорение абсолютного изменения, тыс. т в год	Темп роста уровня, в % к предыдущему	Темп роста уровня, в % к начальному
0 (начальный)	100	—	—	—	100
1	112	12	—	112	112
2	128	16	4	114,3	128
3	148	20	4	115,6	148
4	172	24	4	116,2	172
5	200	28	4	116,3	200
6	232	32	4	116,0	232

Если абсолютное изменение отрицательно, его следует называть абсолютным сокращением. Абсолютное изменение имеет ту же единицу измерения, что и уровни ряда, с добавлением единицы времени, за которую определено изменение: 22 тыс. т в год (или 1,83 тыс. т в месяц, или 110 тыс. т в пятилетие). Без указания единицы времени абсолютный прирост нельзя правильно интерпретировать.

В табл. 12.2 абсолютное изменение уровня не является константой тенденции. Оно со временем возрастает, т.е. уровни ряда изменяются с ускорением. *Ускорение* — это разность между абсолютным изменением за данный период и абсолютным изменением за предыдущий период одинаковой длительности:

$$\Delta'_i = \Delta_i - \Delta_{i-1}.$$

Показатель ускорения применяется только в цепном варианте, но не в базисном. Отрицательная величина ускорения говорит о замедлении роста или об ускорении снижения уровней ряда.

Как видно из табл. 12.2, ускорение является константой тенденции данного ряда, что свидетельствует о параболической форме этой тенденции. Ее уравнение имеет вид:

$$\hat{y}_i = a + bt_i + ct_i^2, \qquad (12.2)$$

где a — уровень ряда в начальный период при $t = 0$;
b — средний по ряду абсолютный прирост;
t_i — номер периода;
c — половина ускорения.

По данным табл. 12.2 имеем:

$$y_i = 100 + 10t_i + 2t_i^2.$$

Показатель ускорения абсолютного изменения уровней выражается в единицах измерения уровня, деленных на квадрат длины периода. В нашем случае ускорение составило 4 тыс. т в год за год или 4 тыс. т год^2. Смысл показателя следующий: объем производства (или добычи угля, руды) имел абсолютный прирост, возрастающий на 4 тыс. т в год ежегодно.

Усвоить рассмотренные показатели поможет следующая аналогия с механическим движением: уровень — это прой-

денный путь, причем начало его отсчета не в нулевой точке. Абсолютный прирост — скорость движения тела, а ускорение абсолютного прироста — ускорение движения. Пройденный путь, считая и тот, который уже был пройден до начала отсчета времени в данной задаче, равен:

$$s = s_0 + v_0 t + \frac{at^2}{2},$$

где s_0 — путь, пройденный до начала отсчета времени;
v_0 — начальная скорость;
t — время, прошедшее с начала его отсчета в задаче;
a — ускорение.

Сравнивая формулу пути с формулой (12.2), видим, что s_0 — аналог свободного члена a; v_0 — аналог абсолютного изменения b; $a/2$ — аналог ускорения прироста c.

Как показано в гл. 3, система показателей должна содержать не только абсолютные, но и относительные статистические показатели.

Относительные показатели динамики необходимы для сравнения разных объектов, особенно если их абсолютные характеристики различны. Предположим, предприятие увеличивало производство аналогичной продукции с тенденцией, выраженной уравнением тренда: $y_i = 20 + 4t + 0{,}5t_i^2$. И абсолютный прирост, и ускорение роста объема продукции во втором предприятии гораздо меньше, чем в первом. Но можно ли ограничиться этими показателями и сделать вывод, что развитие второго предприятия более медленное, чем первого? Меньший уровень еще не есть замедленный темп развития, и это покажет относительная характеристика тенденции динамики — темп роста.

Темп роста — это отношение сравниваемого уровня (более позднего) к уровню принятому за базу сравнения (более раннему). Темп роста исчисляется в цепном варианте к предыдущему уровню или в *базисном* варианте — к одному и тому же, обычно начальному уровню (12.3). Он говорит о том, сколько процентов составляет сравниваемый уровень по отношению к уровню, принятому за базу, или во сколько раз сравниваемый уровень больше уровня, принятого за базу. При этом если уровни снижаются со временем, то сказать, что последующий уровень «больше в 0,33 раза», или составляет 33,3% базового уровня, это, разумеется, означает, что уровень уменьшился в 3 раза. Но сказать, что «уровень меньше в 0,33 раза», это неверно. Темп изменения в разах всегда говорит о том, во сколько раз сравниваемый уровень больше.

Теперь можно сказать, что относительная характеристика роста объема продукции на первом предприятии в среднем за год близка к 115% (рост приблизительно 15% за год), и за шесть лет продукция увеличилась в 2,32 раза. По второму предприятию, в чем может убедиться читатель, вычислив также шесть уровней параболического тренда, в среднем за год объем продукции возрастал примерно на 20%, а за шесть лет — в 3,1 раза. Следовательно, в относительном выражении объем продукции на втором предприятии возрастал быстрее. Только в сочетании абсолютных и относительных характеристик динамики можно правильно отразить процесс развития совокупности (объекта).

Рассмотрим связь абсолютных и относительных показателей динамики. Обозначим темп изменения через k. Тогда имеем:

$$\text{цепной темп роста: } k_{i/i-1} = \frac{y_i}{y_{i-1}}; \qquad (12.3)$$

$$\text{базисный темп роста: } k_{i/0} = \frac{y_i}{y_0}.$$

Если сравниваемый уровень y выразить через уровень предыдущего года плюс прирост или через уровень базисного года плюс базисное абсолютное изменение, получим:

$$k_{i/i-1} = \frac{y_{i-1} + \Delta_i}{y_{i-1}} = 1 + \frac{\Delta_i}{y_{i-1}} \text{ или } 100\% + \frac{\Delta_i}{y_{i-1}} \cdot 100;$$
$$k_{i/0} = \frac{y_0 + \Delta_{0i}}{y_0} = 1 + \frac{\Delta_{0i}}{y_0} \text{ или } 100\% + \frac{\Delta_{0i}}{y_0} \cdot 100. \qquad (12.4)$$

Величину $\Delta_i : y_{i-1}$ или $\Delta_{0i} : y_0$, т.е. отношение абсолютного изменения к предыдущему или базисному уровню, часто называют *относительным приростом* (относительным измене-

нием) или же *темпом прироста*. Он равен $k - 1$ или $k - 100\%$. Темп прироста может иметь как положительные, так и отрицательные значения. Например, финансовый результат от реализации продукции предприятием может быть прибылью (+), а может быть убытком (–), тогда темп изменения и темп прироста применять нельзя.

В этом случае такие показатели теряют смысл и не имеют экономической интерпретации. Сохраняют смысл только абсолютные показатели динамики.

Рассмотрим соотношение между цепными и базисными показателями на примере данных табл. 12.2.

1. Сумма цепных абсолютных изменений равна базисному абсолютному изменению:

$$\sum_{(i)} \Delta_{i(\text{цепн})} = \Delta_{i(\text{баз})}. \qquad (12.5)$$

По данным табл. 12.2 получим:

$$12 + 16 + 20 + 24 + 28 + 32 = 232 - 100 = 132.$$

2. Произведение цепных темпов изменения равно базисному темпу изменения:

$$\prod_{(i)} k_{i(\text{цепн})} = k_{i(\text{баз})}. \qquad (12.6)$$

По данным табл. 12.2 получим:

$$1{,}12 \cdot 1{,}143 \cdot 1{,}156 \cdot 1{,}162 \cdot 1{,}163 \cdot 1{,}16 = 2{,}32.$$

Сумма цепных темпов прироста не равна базисному темпу прироста:

$$12 + 13{,}3 + 15{,}6 + 16{,}2 + 16{,}3 + 16 \neq 132 \text{ (в процентах)}.$$

Значения цепных темпов прироста, рассчитанных каждый к своей базе, различаются не только числом процентов, но и величиной абсолютного изменения, составляющей каждый процент. Поэтому складывать или вычитать цепные темпы прироста нельзя. Абсолютное значение 1% прироста равно сотой части предыдущего, или базисного, уровня.

12.4. Особенности показателей динамики для рядов, состоящих из относительных уровней

Уровнями динамического ряда могут быть не только абсолютные показатели. Ряды динамики могут отражать развитие структуры совокупности, вариации признака в совокупности, взаимосвязи между признаками, соотношения значений признака для разных объектов. В этих случаях уровни динамического ряда сами являются относительными показателями, нередко выражаются в процентах. Следовательно, абсолютные изменения (и ускорения) тоже окажутся относительными величинами, могут быть выражены в процентах, как и темпы изменения, и относительные приросты. Все это создает нередко путаницу в интерпретации и использовании показателей динамики в печати и даже в специальной экономической литературе.

Пример. В США с конца XIX в. для группы ведущих акционерных компаний исчисляется так называемый *индекс Доу-Джонса — арифметическая средняя величина котировок акций на фондовых биржах*. Этот показатель характеризует хозяйственную конъюнктуру: если индекс Доу-Джонса повышается, т.е. растет относительная цена акций, значит, вкладчики капитала рассчитывают получить по акциям бóльший дивиденд (распределяемая часть прибыли). Это говорит о росте деловой активности. Падение индекса Доу-Джонса свидетельствует о снижении деловой активности в стране. Величина этого показателя — отношение в процентах цены акций на бирже к их номиналу (первоначальной цене при выпуске). Это отношение зависит не только от колебаний деловой активности, но имеет также общую тенденцию роста ввиду инфляции — падения покупательной силы доллара США. С начала века этот рост значителен, поэтому в наше время индекс Доу-Джонса составляет более 8000% (акция, когда-то выпущенная на сумму 100 долл., теперь стоит более 8000 долл.).

15 августа 1997 г. индекс Доу-Джонса упал с 7942 до 7694%. Абсолютное изменение индекса составило 248, конечно, не процентов, а пунктов, ведь снизиться больше, чем на 100%, величина не может. Падение даже на 60% создало бы впечатление о полном крахе экономики США.

На деле темп падения индекса Доу-Джонса составлял: 7694 : 7942 − 100% = −3,1%. С 9 по 13 февраля 1998 г. индекс Доу-Джонса вырос с 8190 до 8370%, или на 180 пунктов. А темп роста в процентах составил: 8370% : 8190 = 1,022, или 102,2%. Аналогичные термины должны применяться к динамике показателей структуры. Например, общее производство электроэнергии в Российской Федерации в 1980 г. составляло 805 млрд кВт · ч, в том числе на атомных электростанциях 54 млрд кВт · ч, т.е. их доля была равна 6,7%. В 1995 г. общее производство электроэнергии составило 860 млрд кВт · ч, в том числе на АЭС 99,5 млрд кВт · ч, или 11,6%. Доля АЭС возросла за 15 лет на: 11,6 − 6,7 = 4,9 процентных пункта. А темп роста доли АЭС составил: 11,6% : 6,7% = 1,73. Доля АЭС возросла на 73%.

Показатели динамики долей имеют еще одну особенность, обусловленную тем, что сумма всех долей в любой период времени равна единице, или 100%. Поэтому изменение, произошедшее с одной из долей неизбежно меняет и доли всех других частей целого, если даже по абсолютной величине эти части не изменились. Казалось бы, это положение очевидно, однако нередко в печати встречаются рассуждения о том, что увеличение доли пшеницы и ячменя среди зерновых культур — это хорошо, но плохо, что уменьшились доли ржи, овса и гречихи. Как будто все доли сразу могут увеличиться!

Если признак варьирует альтернативно, то увеличение доли одной группы равно уменьшению доли другой группы в пунктах, но темпы изменения долей в процентах при этом могут сильно различаться. Темп больше у той доли, которая в базисном периоде была меньше — темп прироста (изменения) понимается по абсолютной величине, по модулю. Например, в 1992 г. оплата труда составила 73,6% всех денежных доходов населения России, а прочие доходы — 26,4%. В 2002 г. оплата труда составила только 66,2% всех денежных доходов населения, а доля прочих доходов возросла до 33,8%. Темп прироста доли прочих доходов составил 128%, т.е. их доля возросла на 28%. Доля же оплаты труда сократилась в относительном выражении на 10,1% [$\frac{66,2}{73,6} - 1 = 0,899 - 1 = -0,101$][1].

[1] Россия в цифрах. 2003. Статистический сборник / Госкомстат России. — М.: Финансы и статистика, 2003. — С. 102—103.

В общем виде темп роста одной из альтернативных долей зависит от темпа роста другой доли и величины этой доли следующим образом:

$$k_2 = \frac{1 - k_1 x_0}{1 - x_0}, \qquad (12.7)$$

где k_2 — темп изменения доли второго альтернативного значения признака;

k_1 — темп роста этой доли;

x_0 — доля в базисном периоде одного из альтернативных значений признака.

Абсолютное изменение долей в пунктах зависит от величины доли и темпа роста таким образом:

$$\Delta_1 = -\Delta_2 = x_0(k_1 - 1) \cdot 100. \qquad (12.8)$$

При наличии в совокупности не двух, а более групп абсолютное изменение каждой из долей в пунктах зависит от доли этой группы в базисном периоде и соотношения темпа роста абсолютной величины объемного признака этой группы со средним темпом роста объемного признака по всей совокупности. Доля i-й группы в сравниваемый (текущий) период определяется как

$$d_{i1} = d_{i0}\frac{k_i}{\bar{k}} = \frac{d_{i0}k_i}{\sum_{i=1}^{m} d_{i0}k_i}, \qquad (12.9)$$

где d_{i1}, d_{i0} — доли i-й группы в базисном и текущем периодах соответственно;

k_i — темп роста объемного признака в i-й группе;

$\bar{k}$ — средний темп роста;

m — число групп.

Рассмотрим распределение занятого населения России по формам собственности (табл. 12.3).

Согласно формуле (12.9) доля работающих в организациях с государственной и муниципальной формами собственности в 1998 г. составила:

Таблица 12.3

Распределение занятого населения в России по формам собственности

Форма собственности	Доля в 1992 г., %	Темп изменения численности в 1998 г. к 1992 г., %
Государственная и муниципальная	68,9	49,06
Частная	18,3	201,27
Общественная	0,8	65,57
Совместная и смешанная	12,0	142,64
Всего занятых	100,0	88,30

Источник. Россия в цифрах. 1999. Статистический сборник. Госкомстат России. — М.: Финансы и статистика, 1999. — С. 80.

$$0{,}689 \cdot \frac{0{,}4906}{0{,}689 \cdot 0{,}4906 + 0{,}183 \cdot 2{,}0127 + 0{,}008 \cdot 0{,}6557 + 0{,}12 \cdot 1{,}4264} =$$

$$= \frac{0{,}3380}{0{,}8830} = 0{,}3828, \text{ или } 38{,}28\%.$$

Доля работающих в частном секторе: $0{,}183\frac{2{,}0127}{0{,}883} =$ $= 0{,}4171$, или 41,71%.

Доля работающих в общественных организациях: $0{,}008 \times \frac{0{,}6557}{0{,}883} = 0{,}0059$, или 0,6%.

Доля работающих в совместных и предприятиях смешанной формы собственности:

$$\frac{0{,}12 \cdot 1{,}4264}{0{,}883} = 0{,}1938, \text{ или } 19{,}38\%.$$

Знаменатели всех дробей — 0,883 — это средний (общий) темп изменения численности всех занятых.

Особенностью показателей динамики относительных величин интенсивности является то, что темпы роста и темпы прироста (или сокращения) прямого и обратного показателей не совпадают.

Пример. Трудоемкость производственной операции на старом станке составляла 10 мин., а производительность труда — 48 операций за смену. После замены станка на новый трудоемкость операции снизилась в 5 раз (до 2 мин.), а производительность возросла в те же 5 раз — до 240 операций за смену. Относительное изменение трудоемкости составило: $(2 - 10) : 10 = -0{,}8$, т.е. трудоемкость снизилась на 80%. Относительное изменение производительности труда составило: $(240 - 48) : 48 = 4$, или 400%, т.е. производительность труда возросла на 300%. Причина заключается в том, что пределом, к которому стремятся по мере прогресса показатели ресурсоотдачи, является бесконечность, а пределом, к которому стремятся обратные им показатели ресурсоемкости, является нуль. Понимание поведения показателей динамики прямых и обратных мер эффективности очень важно для экономиста и статистика.

По мере приближения относительного показателя к пределу одно и то же абсолютное изменение в пунктах приобретает иное качественное содержание. Например, если показатель тесноты связи — коэффициент детерминации — возрос с 40 до 65% (на 25 пунктов), то система факторов в регрессионном уравнении как была, так и осталась неполной, хорошей модели не получено. Но если после изменения состава факторов коэффициент детерминации возрос с 65 до 90% — на те же 25 пунктов, это изменение имеет другое качественное содержание: получена хорошая регрессионная модель, в основном объясняющая вариацию результативного признака достаточно полной системой факторов.

12.5. Средние показатели тенденции динамики

Средние показатели динамики — средний уровень ряда, средние абсолютные изменения и ускорения, средние темпы роста — характеризуют тенденцию. Они необходимы при обобщении характеристик тенденции за длительный период, по различным периодам и незаменимы при сравнении развития за неодинаковые по длительности отрезки времени, при выборе аналитического выражения тренда. При наличии в динамическом ряду существенных колебаний уровней определение средних показателей тенденции требует использования специальных методов статистики, которые рассматриваются в следующих разделах. В данном разделе рассматриваются только форма, математические свойства средних пока-

зателей динамики и простейшие приемы их вычисления, применимые на практике к рядам со слабой колеблемостью.

Средний уровень интервального ряда динамики определяется как простая арифметическая средняя из уровней за равные промежутки времени:

$$\bar{y} = \frac{\sum_{(i)}^{n} y_i}{n}, \qquad (12.10)$$

или как взвешенная арифметическая средняя из уровней за неравные промежутки времени, длительность которых и является весами.

Пример. По данным табл. 12.1 определим среднегодовые уровни урожайности картофеля по пяти-шестилетиям.

Период	Среднегодовые уровни, ц/га
Пятилетие 1989—1993	157,0
Шестилетие 1994—1999	184,8

Средние уровни принято условно относить к середине интервала времени, т.е. для пятилетия 1989—1993 гг. — к 1991 г., для шестилетия 1994—1999 гг. — к середине между 1996 и 1997 гг., т.е. к 1996,5 (к 01.07.1996 г.). При неравных промежутках времени, если, например, с 1-го по 18-е число месяца на предприятии работали 45 человек, с 18-го по 27-е — 48 человек, с 28-го по 31-е число — 54 человека, среднесписочное число работников за месяц составит:

$$\bar{y} = \frac{45 \cdot 18 + 48 \cdot 9 + 54 \cdot 4}{31} = 47{,}03 \text{ чел.}$$

В моментном ряду смысл среднего уровня в том, что он характеризует уже не состояние на отдельный момент, а состояние между начальным и конечным моментом учета. Из этого следует, что роль уровней, относящихся к начальному и конечному моментам, существенно иная, чем роль уровней на момент внутри изучаемого отрезка времени. Начальный и конечный уровни находятся на границе изучаемого интервала, они наполовину относятся к предыдущему и последующему интервалам и лишь наполовину к изучаемому. Уровни, относящиеся к моментам внутри осредняемого интервала, целиком относятся только к нему. Отсюда получаем особую форму средней арифметической величины, называемую *хронологической средней:*

$$\bar{y}_{\text{хрон}} = \left(\frac{y_1}{2} + \sum_{i=2}^{n-1} y_i + \frac{y_n}{2}\right) : (n-1). \qquad (12.11)$$

Методика вычисления среднего уровня моментного ряда при неравных промежутках между моментами является спорной и здесь не рассматривается.

Если известны точные даты изменения уровней моментного ряда, то средний уровень определяется как

$$\bar{y} = \frac{\sum_{(i)} y_i t_i}{\sum t_i}, \qquad (12.12)$$

где t_i — время, в течение которого сохранялся уровень.

Средний абсолютный прирост (абсолютное изменение) определяется как простая арифметическая средняя из абсолютных изменений за равные промежутки времени (цепных абсолютных изменений) или как частное от деления величины базисного абсолютного изменения на число осредняемых отрезков времени от базисного до сравниваемого периода:

$$\bar{\Delta} = \frac{\sum \Delta_i}{n} = \frac{y_n - y_0}{n}. \qquad (12.13)$$

Как уже сказано выше, при наличии существенной колеблемости уровней средний абсолютный прирост (изменение), как и средний темп, следует вычислять, отделив сначала тренд от колебаний (соответствующая методика будет изложена ниже). Прямое определение среднего абсолютного прироста по крайним уровням ряда допустимо, если нет существенных колебаний уровней. Например, добыча угля в России довольно равномерно снижалась с 337 млн т в 1992 г. до 232 млн т в 1998 г.[1]

[1] Россия в цифрах. 1996. Статистический сборник / Госкомстат России. — М.: Финансы и статистика, 1996. — С. 297.

По формуле (12.13) среднее годовое сокращение добычи угля составило: $\bar{\Delta} = \frac{232 - 337}{6} = -17{,}5$ млн т в год. Итак, добыча угля в период 1992—1998 гг. в среднем за год снижалась на 17,5 млн т в год, или на 1,46 млн т в месяц.

Для правильной интерпретации показатель среднего абсолютного изменения должен сопровождаться указанием двух единиц времени: 1) время, за которое он вычислен, к которому относится и которое он характеризует (в нашем примере это 6 лет — 1992—1998 гг.); 2) время, за которое показатель рассчитан, время, входящее в его единицу измерения, — 1 год. Можно рассчитать среднемесячный прирост за пятилетие, среднесуточное изменение за год, за месяц, за квартал.

Среднее ускорение абсолютного изменения применяется реже. Для его надежного расчета даже при слабых колебаниях уровней требуется использовать методику аналитического выравнивания по параболе 2-го порядка (см. подразд. 12.5 и 12.6). Не рекомендуется измерять среднее ускорение без абстрагирования от колебаний уровней. Для более грубого, приближенного расчета среднего ускорения можно воспользоваться средними годовыми уровнями, сглаживающими колебания. Например, среднегодовое производство мяса в Российской Федерации составляло:

Годы	Среднегодовое производство мяса, млн т
1976—1980	7,40
1981—1985	8,09
1986—1990	9,68

Абсолютный прирост за второе пятилетие в сравнении с первым составил 0,69 млн т, за третье в сравнении со вторым — 1,59 млн т. Следовательно, ускорение в третьем пятилетии по сравнению со вторым составило: 1,59 − 0,69 = 0,90 млн т в год за пять лет, а среднегодовое ускорение прироста равно: 0,90 : 5 = 0,18 млн т в год за год. Среднее ускорение требует указания трех единиц времени, хотя, как правило, две из них одинаковы: период, на который рассчитан прирост, и время, на которое рассчитано ускорение.

Средний темп изменения определяется наиболее точно при аналитическом выравнивании динамического ряда по экспоненте (см. подразд. 12.5 и 12.6). Если можно пренебречь колеблемостью, то средний темп определяют как *геометрическую среднюю* (см. гл. 5) из цепных темпов роста за n лет или из общего (базисного) темпа роста за n лет:

$$\bar{k} = \sqrt[n]{\prod_{i=1}^{n} k_i} = \sqrt[n]{\frac{y_n}{y_0}}. \qquad (12.14)$$

Например, стоимость потребительской корзины за год в результате инфляции возросла в 6 раз. Каков средний месячный темп инфляции?

$$\bar{k} = \sqrt[12]{6} = 1{,}16, \text{ или } 116\%,$$

т.е. в среднем за месяц цены росли на 16% к уровню предыдущего месяца.

Средний темп роста, так же как и средний прирост, следует сопровождать указанием двух единиц времени: 1) периода, который им характеризуется; 2) периода, на который рассчитан темп. Например, среднегодовой темп за последнее десятилетие; среднемесячный темп за полугодие и т.п.

Если исходной информацией служат темпы прироста и нужно вычислить их среднегодовую величину, то предварительно следует все темпы прироста превратить в темпы роста, прибавив 1, или 100%, вычислить их среднюю геометрическую и затем снова вычесть 1, или 100%. Интересно, что ввиду асимметрии темпа прироста и темпа сокращения при равных их величинах общий темп прироста всегда отрицателен. Так, если за первый год объем производства вырос на 20%, а за второй снизился на 20% (темпы цепные), то за два года имеем:

средний темп роста: $\bar{k} = \sqrt{1{,}2 \cdot 0{,}8} = 0{,}9798$, или 97,98%;
средний темп прироста: $\bar{k} - 1 = -0{,}0202$, или −2,02%.

Как отмечалось в гл. 5, применяя для вычисления среднего темпа среднюю геометрическую, мы опираемся на соблюдение фактического отношения конечного уровня к начальному при замене фактических темпов на средние. В практических задачах может потребоваться вычисление среднего

уровня при условии соблюдения отношения суммы уровней за период к уровню, принятому за базу. Например, если общий выпуск продукции за пятилетие должен составить 800% к базисному (среднегодовому за предыдущие 5 лет выпуску), или, что то же самое, среднегодовой уровень должен составить 160% к базовому уровню, каков должен быть среднегодовой темп роста выпуска продукции? В 1974 г. украинские статистики А. и И. Соляники предложили следующую приближенную формулу для среднего темпа роста, удовлетворяющую этому условию:

$$\overline{k}_{\text{пар}} = 1 + \frac{-3}{2(m-1)} + \sqrt{\frac{9}{4(m-1)^2} + \frac{6}{m(m^2-1)}\left(\frac{\sum_{i=1}^{m} y_i}{y_0} - m\right)}, \quad (12.15)$$

где m — число суммируемых уровней;

y_0 — базисный уровень.

Темп роста данного вида называют параболическим (отсюда обозначение $\overline{k}_{\text{пар}}$), так как он вычисляется по уравнению параболы порядка m. При $m = 5$ имеем:

$$\overline{k}_{\text{пар}} = 1 - \frac{3}{8} + \sqrt{\frac{9}{64} + \frac{1}{20}\left(\frac{\sum y_i}{y_0} - 5\right)} = 1 - 0{,}375 +$$

$$+ \sqrt{0{,}1406 + 0{,}05(8-5)} = 1{,}16407, \text{ или } 116{,}4\%.$$

Расчет по этому среднегодовому темпу дает сумму выпуска за 5 лет в 8,069 раза больше базисной, т.е. приближение хорошее. В общем виде проблема параболических темпов исследована саратовским статистиком Л. С. Казинцом в книге «Темпы роста и абсолютные приросты» (М.: Статистика, 1975). Им составлены таблицы, с помощью которых, зная отношение суммы уровней к базисному уровню и число суммируемых уровней m, можно получить $\overline{k}_{\text{пар}}$. Таблица Л. С. Казинца рассчитана на основе нахождения корней уравнения:

$$\overline{k} + \overline{k}^2 + \overline{k}^3 + \ldots + \overline{k}^n = \sum_{1}^{n} y_i : y_0.$$

Для нашего примера таблица Л. С. Казинца дает среднегодовой темп роста 116,1% и сумму выпуска в 8,00016 раза больше базисной.

Если необходимо определить средний темп изменения исходя из заданной на n периодов суммы абсолютных изменений, то следует использовать формулу:

$$\overline{k} = \sqrt[n]{\frac{y_n}{y_0}} = \sqrt[n]{\frac{y_0 + \sum_{i=1}^{n} \Delta_i}{y_0}} = \sqrt[n]{1 + \frac{\sum_{i=1}^{n} \Delta_i}{y_0}}, \quad (12.16)$$

где y_0 — базисный уровень;

$\sum_{i=1}^{n} \Delta_i$ — сумма абсолютных изменений;

$y_n = y_0 + \sum_{i=1}^{n} \Delta_i$ — конечный уровень.

Рассмотрим применение этой формулы. В 1995 г. в России добыто 262 млн т угля. Каким должен быть среднегодовой темп роста добычи, чтобы к 2000 г. (за 5 лет) сумма абсолютных приростов достигла 150 млн т? Или иначе — чтобы к 2000 г. добыча достигла 412 млн т? Делим 150 млн т на базисный уровень 262 млн т, получаем 57,25% прироста, далее по формуле (12.16) вычисляем среднегодовой темп: $\overline{k} = \sqrt[5]{1{,}5725} = 1{,}09476$, или 109,476%.

Вычисления проверим в табл. 12.4.

Таблица 12.4

Реконструкция ряда динамики по конечному уровню

Год	Добыча, млн т	Абсолютный прирост, млн т/год
1995	262	—
1996	262 · 1,09476 = 286,8	24,8
1997	286,8 · 1,09476 = 314,0	27,2
1998	314,0 · 1,09476 = 343,8	29,8
1999	343,8 · 1,09476 = 376,3	32,5
2000	376,3 · 1,09476 = 412,0	35,7
Итого	1732,9	150,0

Интересную задачу представляет определение срока, за который ряд с бóльшим средним показателем динамики, но меньшим начальным уровнем догонит другой ряд с бóльшим *начальным уровнем, но меньшим показателем динамики.*

Если в качестве показателя динамики берутся средние абсолютные приросты, то задача решается просто. Пусть имеется первый ряд с базисным уровнем y_{01} и средним абсолютным приростом $\bar{\Delta}_1$ и второй ряд с соответствующими показателями y_{02} и $\bar{\Delta}_2$; причем $y_{02} > y_{01}$, $\bar{\Delta}_2 < \bar{\Delta}_1$. Тогда уровень первого ряда сравняется с уровнем второго ряда через

$$\frac{y_{02} - y_{01}}{\bar{\Delta}_1 - \bar{\Delta}_2} \text{ лет.}$$

Та же задача может быть решена на основе ускорений. Имеем первый ряд с базисным уровнем y_{01}, базисным абсолютным изменением a_{01} и средним ускорением b_1; второй ряд — с показателями y_{02}, a_{02}, b_2. При каком числе n периодов (лет) после базисного уровни рядов сравняются?

Тенденции рядов параболические:

$$y_{n1} = y_{01} + a_{01n} + \frac{b_1 n^2}{2};$$

$$y_{n2} = y_{02} + a_{02n} + \frac{b_2 n^2}{2}.$$

Приравняв правые части уравнений, получим:

$$y_{02} + a_{02n} + \frac{b_2 n^2}{2} = y_{01} + a_{01n} + \frac{b_1 n^2}{2}$$

или

$$(y_{02} - y_{01}) + n(a_{02} - a_{01}) + n^2\left(\frac{b_2 - b_1}{2}\right) = 0. \qquad (12.17)$$

Искомый срок n является корнем этого квадратного уравнения. Если, например, имеем:

$$y_{01} = 500; \quad a_{01} = 40; \quad b_1 = 2;$$
$$y_{02} = 300; \quad a_{02} = 26; \quad b_2 = 3,$$

то

$$200 + (-14)n + 0{,}5n^2 = 0.$$

Откуда:

$$n = \frac{14 \pm \sqrt{196 + 400}}{2 \cdot 0{,}5} = 14 \pm 24{,}4.$$

Второй ряд догонит первый по уровню через 38,4 года; в прошлом уровни рядов были одинаковы 10,4 года назад. Будущие равные уровни составляют 3510, а прошлые равны 192.

Если мы хотим найти срок n, через который уровни рядов сравняются, то эту задачу можно решить и на основе средних *темпов динамики.*

Имеем:

$$y_{02} \cdot \bar{k}_2^n = y_{01} \bar{k}_1^n.$$

Логарифмируя это равенство, получаем:

$$n\log\bar{k}_2 + \log y_{02} = n\log\bar{k}_1 + \log y_{01}.$$

Откуда:

$$n(\log\bar{k}_2 - \log\bar{k}_1) = \log y_{01} - \log y_{02}, \qquad (12.18)$$

$$n = \frac{\log y_{01} - \log y_{02}}{\log\bar{k}_2 - \log\bar{k}_1},$$

т.е. искомый срок равен частному от деления разности логарифмов уровней рядов в базисном периоде на разность логарифмов темпов изменения, только переставленных при вычитании. Обычно и в числителе, и в знаменателе из большего логарифма вычитается меньший. Например, первый ряд имеет: $y_{0_1} = 300$; $k_1 = 1{,}09$; второй ряд имеет: $y_{0_2} = 100$; $k_2 = 1{,}2$. Тогда:

$$n = \frac{\ln 300 - \ln 100}{\ln 1{,}2 - \ln 1{,}09} = \frac{5{,}70382 - 4{,}60517}{0{,}18232 - 0{,}08618} = 11{,}43.$$

Через 11,43 года уровень второго ряда сравняется с первым при сохранении экспоненциальных трендов обоих рядов.

12.6. Методы выявления типа тенденции динамики

Прежде чем применить методы математического анализа для вычисления параметров уравнения тренда, необходимо выявить тип тенденции, а эта задача не является чисто математической. Наличие колебаний уровней крайне усложняет выявление типа тенденции и требует всестороннего подхода к этой проблеме, качественного изучения характера развития объекта. При этом нужно дать ответы на такие вопросы:

1. Были ли условия для развития объекта достаточно однородными в изучаемый период?
2. Каков характер действия основных факторов развития?
3. Не произошло ли качественное, существенное изменение условий развития объекта внутри изучаемого периода времени?

Если, например, часть периода предприятие работало по старой технологии, а затем произошло техническое перевооружение — введены новые цехи, поточные линии, то единой тенденции показателей за весь период не будет, скорее всего нужна «периодизация» ряда, т.е. его дробление на отдельные подпериоды: до реконструкции, во время таковой (если она длительна) и после освоения новой технологии.

Чем крупнее изучаемая система, чем больше факторов влияют на динамику изучаемого признака, тем реже возможны резкие, скачкообразные изменения в ряду динамики (не колебания, а именно изменения в тенденции). Большие и сложные системы обладают значительной инерцией, и для скачкообразного, резкого изменения тенденции такой системы требуются большие затраты ресурсов, которые общество выделить не в состоянии. Поэтому такое коренное изменение в экономике, как переход от командно-административного планирования хозяйства к рыночной регулируемой экономике, в масштабе нашей страны неизбежно займет достаточно большое время, за которое сформируются новые тенденции народно-хозяйственных показателей. Для того чтобы разглядеть эти новые тенденции, понадобится время.

Напротив, в масштабе отдельных предприятий вполне возможны резкие изменения, переходы от одной тенденции к другой.

Рассмотрим некоторые основные типы уравнений тренда, выражающие те или иные качественные свойства развития.

1. Линейная форма тренда:

$$\hat{y} = a + bt, \tag{12.19}$$

где $\hat{y}$ — уровни, освобожденные от колебаний, выравненные по прямой;
a — начальный уровень тренда в момент или период, принятый за начало отсчета времени t;
b — среднегодовой абсолютный прирост (среднее изменение за единицу времени); константа тренда.

Линейный тренд хорошо отражает тенденцию изменений при действии множества разнообразных факторов, изменяющихся различным образом по разным закономерностям. Равнодействующая этих факторов при взаимопогашении особенностей отдельных факторов (ускорение, замедление, нелинейность) часто выражается в примерно постоянной абсолютной скорости изменения, т.е. в прямолинейном тренде. Таковы, например, тенденции динамики урожайности для масштаба области, республики, крупного региона, страны в целом.

2. Параболическая форма тренда:

$$\hat{y} = a + bt + ct^2, \tag{12.20}$$

где c — квадратический параметр, равный половине ускорения; константа параболического тренда. Остальные обозначения прежние.

Параболическая форма тренда выражает ускоренное или замедленное изменение уровней ряда с постоянным ускорением. Такой характер развития можно ожидать при наличии важных факторов прогрессивного развития (прогрессирующее поступление нового высокопроизводительного оборудования, увеличение среднесуточного прироста живого веса поросят с возрастом и т.п.). Ускоренное возрастание может происходить в период после снятия каких-то сдерживающих развитие преград — ограничений в распределении дохода, в уровне оплаты труда, при повышении цены на дефицитную продукцию.

Параболическая форма тренда с отрицательным ускорением ($c < 0$) приводит со временем не только к приостановке роста уровня, но и к его снижению со все большей скоростью. Такой характер развития может быть свойствен производству

устаревшей продукции, ликвидируемой отрасли сельского хозяйства на предприятии (ферме) и т.п.

Парабола 2-го порядка (квадратическая) имеет либо максимум (если $c < 0$ и $b > 0$), либо минимум ($b < 0$, $c > 0$). Для нахождения экстремума производную параболы по времени t следует приравнять нулю и решить полученное уравнение относительно t. Например, если население города (тыс. чел.) возрастает по параболе

$$y = 1800 + 80t - 2t^2,$$

то производная по времени $\partial f/\partial t$ будет иметь вид: $80 - 4t = 0$, откуда: $t = 20$. Максимум населения будет достигнут через 20 лет после начала отсчета времени, и это максимальное население составит:

$$\hat{y}_{max} = 1800 + 80 \cdot 20 - 2 \cdot 20^2 = 2600 \text{ тыс. чел.}$$

3. Экспоненциальная форма тренда:

$$\tilde{y} = ak^t, \qquad (12.21)$$

где k — темп изменения в разах;
a — константа тренда.

Если $k > 1$, экспоненциальный тренд выражает тенденцию ускоренного и все более ускоряющегося возрастания уровней. Такой характер свойствен, например, размножению организмов при отсутствии ограничения со стороны среды: сорняков, хищников, вирусных заболеваний. При росте по экспоненте абсолютный прирост пропорционален достигнутому уровню. Так росло население Земли в эпоху «демографического взрыва» в XX столетии; сейчас этот период заканчивается и темп роста населения стал уменьшаться. Если бы он остался на уровне 1960—1970 гг., т.е. около 2% прироста в год от 1985 г., когда население составляло 5 млрд чел., то к 2500 г. население Земли достигло бы уровня: $5 \text{ млрд} \cdot 1{,}02^{515} =$ $= 134$ трлн 286 млрд чел.; на 1 человека приходилось бы примерно 1 м^2 всей площади суши. Ясно, что рост любого объекта по экспоненциальному закону может продолжаться только небольшой исторический период, поскольку любой процесс развития всегда встретит ограничения.

При $k < 1$ экспоненциальный тренд означает тенденцию постоянно все более замедляющегося снижения уровней динамического ряда. Такая тенденция может быть присуща динамике трудоемкости продукции, удельных затрат топлива, металла на единицу полезного эффекта (на 1 кВт · ч, на 1 м^2 жилой площади и т.д.) при технологическом прогрессе. Экстремальных точек экспонента не имеет.

4. Логарифмическая форма тренда:

$$\hat{y} = a + b \log t. \qquad (12.22)$$

Логарифмический тренд пригоден для отображения тенденции замедляющегося роста уровней при отсутствии предельно возможного значения. Замедление роста становится все меньше и меньше, и при достаточно большом t логарифмическая кривая становится малоотличимой от прямой линии. Логарифмический тренд пригоден для отображения роста спортивных достижений (чем они выше, тем труднее их улучшать), роста производительности агрегата по мере его освоения и совершенствования, повышения продуктивности скота или эффективности системы при ее совершенствовании без качественных, коренных преобразований. Экстремума логарифмическая кривая не имеет.

5. Тренд в форме степенной кривой:

$$\hat{y} = at^b, \qquad (12.23)$$

где b — константа тренда.

При $b = 1$ имеем линейный тренд, $b = 2$ — параболический и т.п. Степенная форма — гибкая, пригодная для отображения изменений с разной мерой пропорциональности изменений во времени. Жестким условием является обязательное прохождение через начало координат: при $t = 0$, $y = 0$. Можно усложнить форму тренда: $\hat{y} = a + t^b$ или $\hat{y} = a + ct^b$, но эти уравнения нельзя логарифмировать, трудно вычислять параметры, и они крайне редко применяются.

6. Гиперболическая форма тренда:

$$\hat{y} = a + \frac{b}{t}. \qquad (12.24)$$

Если $b > 0$, гиперболический тренд соответствует тенденции замедляющегося снижения уровня, стремящегося к пределу a. Если $b < 0$, тренд выражает тенденцию замедляющегося

роста уровней, стремящихся в пределе к a. Следовательно, гиперболическая форма тренда подходит для отображения тенденции, процессов, ограниченных предельным значением уровня (предельным коэффициентом полезного действия двигателя, пределом 100%-ной грамотности населения и т.п.).

7. Логистическая форма тренда:

$$\hat{y} = \frac{y_{max} - y_{min}}{e^{a+bt} + 1} + y_{min}, \qquad (12.25)$$

где y_{max} и y_{min} — максимальное и минимальное из возможных значений уровня;
e — основание натуральных логарифмов;
a, b — параметры тренда.

Логистическая кривая имеет форму латинской буквы s, положенной на бок, отчего еще называется *s-образной кривой*. Она имеет два перегиба: от ускоряющегося роста к равномерному (вогнутость) и от равномерного роста посреди периода развития в течение длительного периода к замедляющемуся росту. Логистическая форма тренда характерна для развития, проходящего все фазы, например процесса насыщения потребителей каким-то новым товаром, скажем телевизорами: сначала медленный, но все ускоряющийся рост доли семей, имеющих телевизор, затем рост равномерный (примерно от 30—40% семей до 70—80%). Затем рост доли семей, имеющих телевизор, замедляется по мере приближения доли к 100%. Если $y_{min} = 0$, $y_{max} = 100\%$, или 1, уравнение упрощается и принимает вид:

$$\hat{y} = \frac{1}{e^{a+bt} + 1}. \qquad (12.26)$$

После теоретического исследования особенностей разных форм тренда необходимо обратиться к фактическому ряду динамики, тем более что далеко не всегда можно надежно установить, какой должна быть форма тренда из чисто теоретических соображений. По фактическому динамическому ряду тип тренда устанавливают на основе графического изображения, путем осреднения показателей динамики, на основе статистической проверки гипотезы о постоянстве параметра тренда.

На рис. 12.1 достаточно хорошо видно, что тренд урожайности выражен прямой линией. Исходный ряд уровней короткий, поэтому на данном примере нельзя использовать другие приемы. Применим их к анализу динамики индекса цен на нетопливные товары развивающихся стран за 1979—1995 гг.[1]

Скользящая пятилетняя средняя, сглаживая колебания отдельных уровней, довольно отчетливо показывает тенденцию равномерного снижения уровней. Если разбить ряд на три части, то средние уровни также подтверждают этот вывод: за 1979—1983 гг. средний уровень равен 112,3; за 1984—1989 гг. — 103,0; за 1990—1995 гг. — 97,0. Существенного различия в величине снижения среднегодовых уровней нет. Оба приема — скользящая средняя и средние уровни по частям ряда — несвободны от субъективных факторов. Можно скользящую среднюю вычислять не за 5 лет, а за 6 или 7; можно иначе разбить ряд — на три части или на другое число частей.

Более обоснованным приемом выявления тренда является проверка статистической гипотезы о постоянстве того или иного показателя динамики[2]. Рассмотрим этот прием по данным табл. 12.5.

В первую очередь проверяется гипотеза о наиболее простой — линейной форме уравнения тренда, т.е. о несущественности различий цепных абсолютных изменений. Имеем 12 абсолютных изменений скользящей средней, которая хотя и сгладила сильные колебания уровней ряда, но, как видим, ее абсолютные изменения далеко не одинаковы. Разбиваем эти 12 цепных приростов на два подпериода: по 6 приростов в каждом и для каждого подпериода вычисляем среднюю $\bar{\Delta}_k$, среднее квадратическое отклонение (СКО) как оценку генерального СКО с учетом потери одной степени свободы вариации s:

$$s_{\Delta_k} = \sqrt{\frac{\sum_{i=1}^{5} (\Delta_i - \bar{\Delta})^2}{6-1}}$$

[1] International Monetary fund // World Economic Outlook. — Washington: D. C., 1996. — P. 68.

[2] Прием предложен М. С. Каяйкиной в статье «Выбор типа линии при аналитическом выравнивании динамических рядов урожайности сельскохозяйственных культур» // Записки ЛСХИ. — Ленинград — Пушкин, 1972. — Т. 196.

Таблица 12.5

Проверка гипотезы о линейном тренде индекса цен (1990 г. = 100 %)

Год	Уровни y_i	Суммы за 5 лет	Суммы скользящие за 5 лет $\bar{y}_i$	Цепные приросты $y_i - y_{i1}$
1979	105	...	...	...
1980	111	...	...	
1981	110	550	110,0	...
1982	106	569	113,8	+3,8
1983	118	571	114,2	+0,4
1984	124	553	110,6	−3,6
1985	113	538	107,6	−3,0
1986	92	529	105,8	−1,8
1987	91	518	103,6	−2,2
1988	109	505	101,0	−2,6
1989	113	507	101,4	+0,4
1990	100	507	101,4	0,0
1991	94	490	98,0	−3,4
1992	91	479	95,8	−2,2
1993	92	485	97,0	+1,2
1994	102	...	...	...
1995	106	...	...	...
	1777	—	—	—

и среднюю ошибку среднего изменения m_{Δ_k} по правилам, рассмотренным в гл. 7:

$$m_{\Delta_k} = \frac{s_{\Delta k}}{\sqrt{6}}.$$

$$\bar{\Delta}_1 = \frac{-6,4}{6} = -1,07; \quad s_1 = \sqrt{\frac{37,82}{6-1}} = 2,75; \quad m_1 = s_1 : \sqrt{6} = 1,123;$$

$$\bar{\Delta}_2 = \frac{-6,6}{6} = -1,1; \quad s_2 = \sqrt{\frac{17,50}{6-1}} = 1,87; \quad m_2 = s_2 : \sqrt{6} = 0,763.$$

Для проверки гипотезы о несущественности различий между средними абсолютными изменениями по подпериодам $\bar{\Delta}_1$, $\bar{\Delta}_2$ М. С. Каяйкина предложила проверять существенность их различий попарно по t-критерию Стьюдента. Затем методика

была дополнена и усовершенствована А. И. Манеллей, предложившим проверять существенность всех различий сразу по критерию Фишера.

Средняя случайная ошибка разностей двух выборочных средних оценок, как показано в гл. 7, есть корень квадратный из суммы квадратов ошибок каждой из выборочных средних, т.е.

$$m_{d(1-2)} = \sqrt{m_1^2 + m_2^2} = \sqrt{1,123^2 + 0,763^2} = 1,356.$$

Критерий Стьюдента для оценки существенности различия двух среднегодовых приростов (изменений) составит:

$$t = \frac{-1,07 - (-1,10)}{1,356} = 0,022.$$

Критическое значение t-критерия при уровне значимости 0,05 и при $(6-1) + (6-1) = 10$ степенях свободы равно 2,23 (табл. П.2 приложения). Фактическое значение намного меньше. Следовательно, вероятность того, что различие среднегодовых приростов в разные подпериоды случайно, превышает 0,05, и гипотеза о равенстве приростов не отклоняется. А значит, тенденцию динамики на всем протяжении ряда можно считать линейной.

Если же гипотеза о линейности отклоняется, по скользящим средним и их цепным приростам вычисляют ускорения приростов и аналогичным методом проверяют существенность различия ускорения в подпериодах. Если несущественно различие ускорений, принимается гипотеза о том, что тренд — парабола 2-го порядка. Если и гипотеза о постоянстве ускорений отклоняется, то по скользящей средней вычисляют цепные темпы роста и проверяют гипотезу об их постоянстве по подпериодам. Подтверждение (неотклонение) этой гипотезы означает принятие гипотезы о том, что тренд экспоненциальный.

Проверка гипотез о других типах тенденций динамики, рассмотренных в подразд. 12.4, сложнее и здесь излагаться не будет. Итак, в нашем примере принято решение считать тренд линейным и следует приступить к вычислению его параметров.

12.7. Методика измерения параметров тренда

Когда тип тренда установлен, необходимо вычислить оптимальные значения параметров тренда исходя из фактических уровней. Для этого обычно используют метод наименьших квадратов. Его значение уже рассмотрено в предыдущих главах учебного пособия, в данном случае оптимизация состоит в минимизации суммы квадратов отклонений фактических уровней ряда от выравненных уровней (от тренда). Для каждого типа тренда МНК дает систему нормальных уравнений, решая которую вычисляют параметры тренда. Рассмотрим лишь три такие системы: для прямой, для параболы 2-го порядка и для экспоненты. Приемы определения параметров других типов тренда рассматриваются в специальных монографиях.

Для *линейного тренда* нормальные уравнения МНК имеют вид:

$$na + b\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i;$$

$$a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i t_i, \qquad (12.27)$$

где n — число уровней ряда;
t_i — номер периода или момента времени;
y_i — уровень исходного ряда динамики.

Систему можно упростить, перенеся начало отсчета времени t_i в середину ряда. Тогда Σt_i (а также суммы всех нечетных степеней t_i) будет равна нулю и система приобретет вид:

$$na = \sum_i y_i, \qquad (12.28)$$

$$b\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i t_i,$$

откуда: $a = \bar{y}$; $b = \dfrac{\sum y_i t_i}{\sum t_i^2}$.

Суммы $\sum y_i t_i$ и $\sum t_i^2$ для значений t_i от $-(n-1)/2$ до $+(n-1)/2$ (табл. 12.5), при этом $\sum t_i^2$ можно вычислить, как $(n^3 - n)/12$.

Нормальные уравнения МНК для параболы 2-го порядка имеют следующий вид:

$$na + b\sum_{(i)} t_i + c\sum_{(i)} t_i^2 = \sum_{(i)} y_i;$$

$$a\sum_{(i)} t_i + b\sum_{(i)} t_i^2 + c\sum_{(i)} t_i^3 = \sum_{(i)} y_i t_i; \qquad (12.29)$$

$$a\sum_{(i)} t_i^2 + b\sum_{(i)} t_i^3 + c\sum_{(i)} t_i^4 = \sum_{(i)} y_i t_i^2.$$

После переноса начала отсчета t в середину ряда имеем:

$$na + c\sum_{(i)} t_i^2 = \sum_{(i)} y_i;$$

$$b\sum_{(i)} t_i^2 = \sum_{(i)} y_i t_i; \qquad (12.30)$$

$$a\sum_{(i)} t_i^2 + c\sum_{(i)} t_i^4 = \sum_{(i)} y_i t_i^2.$$

Суммы включают значения t_i от $-(n-1):2$ до $+(n-1):2$, при этом сумма биквадратов может быть вычислена по формуле

$$\sum t^4 = \frac{3n^5 - 10n^3 + 7n}{240}.$$

Нормальные уравнения МНК для экспоненты имеют следующий вид:

$$n \ln a + \ln k \sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \ln y_i;$$

$$\ln a \sum_{i=1}^{n} t_i + \ln k \sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} \ln y_i t_i. \qquad (12.31)$$

После переноса начала отсчета t_i в середину ряда получим:

$$n \ln a = \sum \ln y_i \text{ откуда: } \ln a = \overline{\ln y_i};$$

$$\ln k \sum t_i^2 = \sum \ln y_i t_i, \text{ откуда: } \ln k = \frac{\sum \ln y_i t_i}{\sum t_i^2}. \quad (12.32)$$

В формуле (12.32) имеем суммирование от $t = -(n-1):2$ до $t = (n-1):2$; в целом формула (12.32) аналогична формуле для линейного тренда (12.28).

По данным табл. 12.1 рассчитаем все три перечисленных тренда для динамического ряда урожайности картофеля с целью их сравнения (табл. 12.6).

Согласно формуле (12.29) параметры линейного тренда равны: $a = 1894/11 = 172{,}2$ ц/га; $b = 486/110 = 4{,}418$ ц/га. Уравнение линейного тренда имеет вид: $\hat{y}_t = 172{,}2 + 4{,}418t$, где $t = 0$ в 1994 г. Это означает, что средний фактический и выравненный уровень, отнесенный к середине периода, т.е. к

Расчет параметров

Год	Уровни y_i	t_i	$y_i t_i$	t_i^2	$y_i t_i^2$	t_i^4
1989	149	−5	−745	25	3725	625
1990	145	−4	−580	16	2320	256
1991	168	−3	−504	9	1512	81
1992	146	−2	−292	4	584	16
1993	177	−1	−177	1	177	1
1994	176	0	0	0	0	0
1995	190	1	190	1	190	1
1996	186	2	372	4	744	16
1997	176	3	528	9	1584	81
1998	211	4	844	16	3376	256
1999	170	5	850	25	4250	625
Σ	1894	0	486	110	18462	1958

1994 г., равен 172 ц/га, а среднегодовой прирост составляет 4,418 ц/га в год.

Параметры параболического тренда согласно (12.22) равны: $b = 4{,}418$; $a = 177{,}75$; $c = -0{,}5571$. Уравнение параболического тренда имеет вид: $\hat{y} = 172{,}75 + 4{,}418t - 0{,}5571t^2$; $t = 0$ в 1994 г. Это означает, что абсолютный прирост урожайности замедляется в среднем на: $2 \cdot 0{,}447$ ц/га в год за год. Сам же абсолютный прирост уже не является константой параболического тренда, а является средней величиной за период. В год, принятый за начало отсчета, т.е. 1994 г., тренд проходит через точку с ординатой 177,75 ц/га. Свободный член параболического тренда не является средним уровнем за период. Параметры экспоненциального тренда вычисляются по формулам (12.31) и (12.32): $\ln a = 56{,}5658/11 = 5{,}1423$; потенцируя, получаем: $a = 171{,}1$; $\ln k = 2{,}853 : 110 = 0{,}025936$; потенцируя, получаем: $k = 1{,}02628$.

Уравнение экспоненциального тренда имеет вид: $\hat{y} = 171{,}1 \cdot 1{,}02628^t$. Это означает, что среднегодовой темп роста урожайности за период составил 102,63%. В точке, принятой за начало отсчета, тренд проходит точку с ординатой 171,1 ц/га.

Таблица 12.6

трендов

$\ln y_i$	$\ln y_i t_i$	Уровни трендов $\hat{y}_i$		
		линейный	параболы 2-го порядка	экспоненты
5,0039	−25,020	150	141,7	150,3
4,9767	−19,907	155	151,2	154,2
5,1240	−15,372	159	159,5	158,3
4,9836	−9,967	163	166,7	162,4
5,1762	−5,176	168	172,8	166,7
5,1705	0	172	177,7	171,1
5,2470	5,247	177	181,6	175,6
5,2257	10,451	181	184,4	180,2
5,1705	15,511	185	186,0	184,9
5,3519	21,407	190	186,5	189,8
5,1358	25,679	194	185,9	194,8
56,5658	2,853	1894	1894	1888,3

Рассчитанные по уравнениям трендов уровни записаны в трех последних графах табл. 12.6. Как видно по этим данным, расчетные значения уровней по всем трем видам трендов различаются ненамного, так как и ускорение параболы, и темп роста экспоненты невелики. Существенное отличие имеет парабола — рост уровней с 1998 г. прекращается, в то время как при линейном тренде уровни растут и далее, а при экспоненте их рост ускоряется. Поэтому для прогнозов эти три тренда неравноправны: при экстраполяции параболы на будущие годы уровни резко разойдутся с прямой и экспонентой, что видно из табл. 12.7. В этой таблице представлена распечатка решения на ПЭВМ по программе «Statgraphics» тех же трендов. Отличие их свободных членов от приведенных выше объясняется тем, что программа нумерует года не от середины, а от начала, так что свободные члены трендов относятся к 1998 г., для которого $t = 0$. Уравнение экспоненты на распечатке составлено в логарифмированном виде. Прогноз сделан на 5 лет вперед, т.е. до 2004 г. При изменении начала координат (отсчета времени) в уравнении параболы меняется и средний абсолютный прирост, параметр b, так как в результате отрицательного ускорения прирост все время сокращается, а его максимум — в начале периода. Константой параболы является только ускорение.

В строке «Data» приводятся уровни исходного ряда; «Forecast summary» означает сводные данные для прогноза. В последних трех строках приведены результаты уравнению прямой, по уравнению параболы и по экспоненте в логарифмическом виде. Графа ME означает среднее расхождение между уровнями исходного ряда и уровнями тренда (выравненными). Для прямой и параболы это расхождение всегда равно нулю. Уровни экспоненты в среднем на 0,48852 ниже уровней исходного ряда. Точное совпадение возможно, если истинный тренд — экспонента; в данном случае совпадения нет, но различие мало. Графа MSE — это дисперсия s^2, мера колеблемости фактических уровней относительно тренда, о чем сказано в подразд. 12.7. Графа MSE — среднее линейное отклонение уровней от тренда по модулю (см. подразд. 5.8); графа MAPE — относительное линейное отклонение в процентах. Здесь они приведены как показатели пригодности выбранного вида тренда. Меньшую дисперсию и модуль отклонения имеет парабола:

Таблица 12.7

Распечатка решения на ПЭВМ по программе «Statgraphics»

Data	149	145	168	146	177	176	190	186	176	211	170
Forecast summary						ME	MSE	MAE	MAPE	MPE	
145,673 + 4,41818 × T						0,00000	173,127	11,2099	6,48645	−0,56996	
131,188 + 1,1035 × T − 0,55711 × T ^2						0,00000	148,918	9,9218	5,79005	−0,47525	
EXP (4,98665 + 0,02594498 × T)						0,48852	178,764	11,5423	6,66106	−0,28789	
Period	Period				Period						
12	13		14		15		16				
198,691	203,109		207,527		211,945		216,364				
184,206	181,382		177,443		172,391		166,224				
199,945	205,202		210,596		216,133		221,815				

она за период 1989—1999 гг. ближе к фактическим уровням. Но выбор типа тренда нельзя сводить лишь к этому критерию. На самом деле замедление прироста есть результат большого отрицательного отклонения, т.е. неурожая в 1999 г.

Вторая половина таблицы — это прогноз уровней урожайности по трем видам трендов на годы; t_i = 12, 13, 14, 15 и 16 от начала отсчета (1989 г.). Прогнозируемые уровни по экспоненте вплоть до 16-го года ненамного выше, чем по прямой. Уровни тренда-параболы снижаются, все более расходясь с другими трендами.

Как видно из табл. 12.6, при вычислении параметров тренда уровни исходного ряда входят с разными весами — значениями t_i и их квадратов. Поэтому влияние колебаний уровней на параметры тренда зависит от того, на какой номер года приходится урожайный либо неурожайный год. Если резкое отклонение приходится на год с нулевым номером ($t_i = 0$), то оно никакого влияния на параметры тренда не окажет, а если попадет на начало и конец ряда, то повлияет сильно. Следовательно, однократное аналитическое выравнивание неполно освобождает параметры тренда от влияния колеблемости, что в нашем примере случилось с параболой. Для дальнейшего исключения искажающего влияния колебаний на параметры тренда следует применить *метод многократного скользящего выравнивания*.

Этот прием состоит в том, что параметры тренда вычисляются не сразу по всему ряду, а скользящим методом, сначала за первые m периодов времени или моментов, затем за период от 2-го до $m + 1$, от 3-го до $(m + 2)$-го уровня и т.п. Если число исходных уровней ряда равно n, а длина каждой скользящей базы расчета параметров равна m, то число таких скользящих баз L составит:

$$L = n + 1 - m.$$

Применение методики скользящего выравнивания можно рассматривать, как видно из приведенных расчетов, только при достаточно большом числе уровней ряда, как правило, 15 и более. Рассмотрим эту методику на примере данных табл. 12.5 — динамики цен на нетопливные товары развивающихся стран,

что опять же дает возможность читателю участвовать в небольшом научном исследовании. На этом же примере продолжим рассмотрение методики прогнозирования в подразд. 12.10.

Если вычислять в нашем ряду параметры по 11-летним периодам (по 11 уровням), то $L = 17 + 1 - 11 = 7$. Смысл многократного скользящего выравнивания в том, что при последовательных сдвигах базы расчета параметров на концах ее и в середине окажутся разные уровни с разными по знаку и величине отклонениями от тренда.

Поэтому при одних сдвигах базы параметры будут завышаться, при других — занижаться, а при последующем усреднении значений параметров по всем сдвигам базы расчета произойдет дальнейшее взаимопогашение искажений параметров тренда колебаниями уровней.

Многократное скользящее выравнивание не только позволяет получить более точную и надежную оценку параметров тренда, но и осуществить контроль правильности выбора типа уравнения тренда. Если окажется, что ведущий параметр тренда, его константа, при расчете по скользящим базам не беспорядочно колеблется, а систематически изменяет свою величину существенным образом, значит, тип тренда был выбран неверно, данный параметр константой не является.

Что касается свободного члена при многократном выравнивании, то нет необходимости и, более того, просто неверно вычислять его величину как среднюю по всем сдвигам базы, потому что при таком способе отдельные уровни исходного ряда входили бы в расчет средней с разными весами и сумма выравненных уровней разошлась бы с суммой членов исходного ряда.

Свободный член тренда — это средняя величина уровня за период при условии отсчета времени от середины периода. При отсчете от начала, если первый уровень $t_i = 1$, свободный член будет равен: $a_0 = \bar{y} - b((n - 1)/2)$. Рекомендуется выбирать длину скользящей базы расчета параметров тренда не менее 9—11 уровней, чтобы в достаточной мере погасить колебания уровней. Если исходный ряд очень длинный, база может составлять до 0,7—0,8 его длины. Для устранения влия-

ния долгопериодических (циклических) колебаний на параметры тренда число сдвигов базы должно быть равно или кратно длине цикла колебаний. Тогда начало и конец базы будут последовательно «пробегать» все фазы цикла и при усреднении параметра по всем сдвигам его искажения от циклических колебаний будут взаимопогашаться. Другой способ — взять длину скользящей базы, равной длине цикла, чтобы начало и конец базы всегда приходились на одну и ту же фазу цикла колебаний.

Поскольку по данным табл. 12.5 уже было установлено, что тренд имеет линейную форму, проводим расчет среднегодового абсолютного прироста, т.е. параметра b уравнения линейного тренда, скользящим способом по 11-летним базам (табл. 12.8).

В этой же таблице приведен расчет данных, необходимых для последующего изучения колеблемости в подразд. 12.7. Остановимся подробнее на методике многократного выравнивания по скользящим базам.

Рассчитаем параметр b по всем базам:

$$b_1 = \frac{-58}{110} = -0{,}527; \quad b_2 = \frac{-120}{110} = -1{,}091; \quad b_3 = \frac{-171}{110} = -1{,}535;$$

$$b_4 = \frac{-226}{110} = -2{,}055; \quad b_5 = \frac{-281}{110} = -2{,}555; \quad b_6 = \frac{-200}{110} = -1{,}818;$$

$$b_7 = \frac{-47}{110} = -0{,}427; \quad \overline{b} = -1{,}433; \quad a = \frac{1777}{17} = 104{,}53.$$

Уравнение тренда: $\hat{y} = 104{,}53 - 1{,}433t$; $t = 0$ в 1987 г.

Итак, индекс цен в среднем за год снижался на 1,433 пункта. Однократное выравнивание по всем 17 уровням может исказить этот параметр, так как начальный уровень содержит значительное отрицательное отклонение, а конечный уровень — положительное. В самом деле, однократное выравнивание дает величину среднегодового снижения индекса цен всего на 0,953 пункта.

Таблица 12.8

Многократное скользящее выравнивание по прямой

Год	Уровни y_i	I база		II база		III база		IV база		V база		VI база		VII база		Тренд, $\hat{y}_i$	Колебания, $y_i - \hat{y}_i = u_i$	u_i^2	$u_i \cdot u_{i+1}$	Поворотные точки
		t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$					
1979	105	−5	−525													116,0	−11,0	121,00	...	
1980	111	−4	−444	−5	−555											114,6	−3,6	12,96	+39,6	
1981	110	−3	−330	−4	−440	−5	−550									113,1	−3,1	9,61	+11,16	•
1982	106	−2	−212	−3	−318	−4	−424	−5	−530							111,7	−5,7	32,49	+17,67	•
1983	118	−1	−118	−2	−236	−3	−354	−4	−472	−5	−590					110,3	+7,7	59,29	−43,89	
1984	124	0	0	−1	−124	−2	−248	−3	−372	−4	−496	−5	−620			108,8	+15,2	231,04	+117,04	•
1985	113	1	113	0	0	−1	−113	−2	−226	−3	−339	−4	−452	−5	−565	107,4	+5,6	31,36	+85,12	
1986	92	2	184	1	92	0	0	1	−92	−2	−184	−3	−276	−4	−368	106,0	−14,0	196,00	−78,40	•
1987	91	3	273	2	182	1	91	0	0	−1	−91	−2	−182	−3	−273	104,5	−13,5	182,25	+189,00	
1988	109	4	436	3	327	2	218	1	109	0	0	−1	−109	−2	−218	103,1	+5,9	34,81	−79,65	
1989	113	5	565	4	452	3	339	2	226	1	113	0	0	−1	−113	101,7	+11,3	127,69	+66,67	•
1990	100			5	500	4	400	3	300	2	200	1	100	0	0	100,2	−0,2	0,04	−2,26	
1991	94					5	470	4	376	3	282	2	188	1	94	98,8	−4,8	23,04	+0,96	
1992	91							5	455	4	364	3	273	2	182	97,4	−6,4	40,96	+30,72	•
1993	92									5	460	4	368	3	276	95,9	−3,9	15,21	+24,96	
1994	102											5	510	4	408	94,5	+7,5	56,25	−29,25	
1995	106													5	530	93,1	+12,9	166,41	+96,75	
Σ	1777	—	−58	—	−120	—	−171	—	−226	—	−281	—	−200	—	−47	1777,1	—	1340,41	+446,20	6

12.8. Методика изучения и показатели колеблемости

Если при изучении и измерении тенденции динамики колебания уровней играли лишь роль помех, «информационного шума», от которого следовало по возможности абстрагироваться, то в дальнейшем сама колеблемость становится предметом статистического исследования. Значение изучения колебаний уровней динамического ряда очевидно: колебания урожайности, продуктивности скота, производства мяса экономически нежелательны, так как потребность в продукции агрокомплекса постоянна. Эти колебания следует уменьшать, применяя прогрессивную технологию и другие меры. Напротив, сезонные колебания объемов производства зимней и летней обуви, одежды, мороженого, зонтиков, коньков необходимы и закономерны, так как спрос на эти товары тоже колеблется по сезонам и равномерное производство требует лишних затрат на хранение запасов. Регулирование рыночной экономики как со стороны государства, так и производителей в значительной мере состоит в регулировании колебаний экономических процессов.

Типы колебаний статистических показателей весьма разнообразны, но все же можно выделить три основных: пилообразную, или маятниковую, колеблемость, циклическую долгопериодическую и случайно распределенную по времени колеблемость. Их свойства и отличия друг от друга хорошо видны при графическом изображении (рис. 12.2).

Пилообразная, или маятниковая, колеблемость состоит в попеременных отклонениях уровней от тренда в одну и в другую сторону. Таковы автоколебания маятника. Подобные автоколебания можно наблюдать в динамике урожайности при невысоком уровне агротехники: высокий урожай при благоприятных условиях погоды выносит из почвы больше питательных веществ, чем образуется естественным путем за почва обедняется, что вызывает снижение следующего урожая ниже тренда, он выносит меньше питательных веществ, чем образуется за год; плодородие возрастает и т.д.

Циклическая долгопериодическая колеблемость свойственна, например, солнечной активности (10—11-летние циклы), а значит, и связанным с ней на Земле процессам — полярным

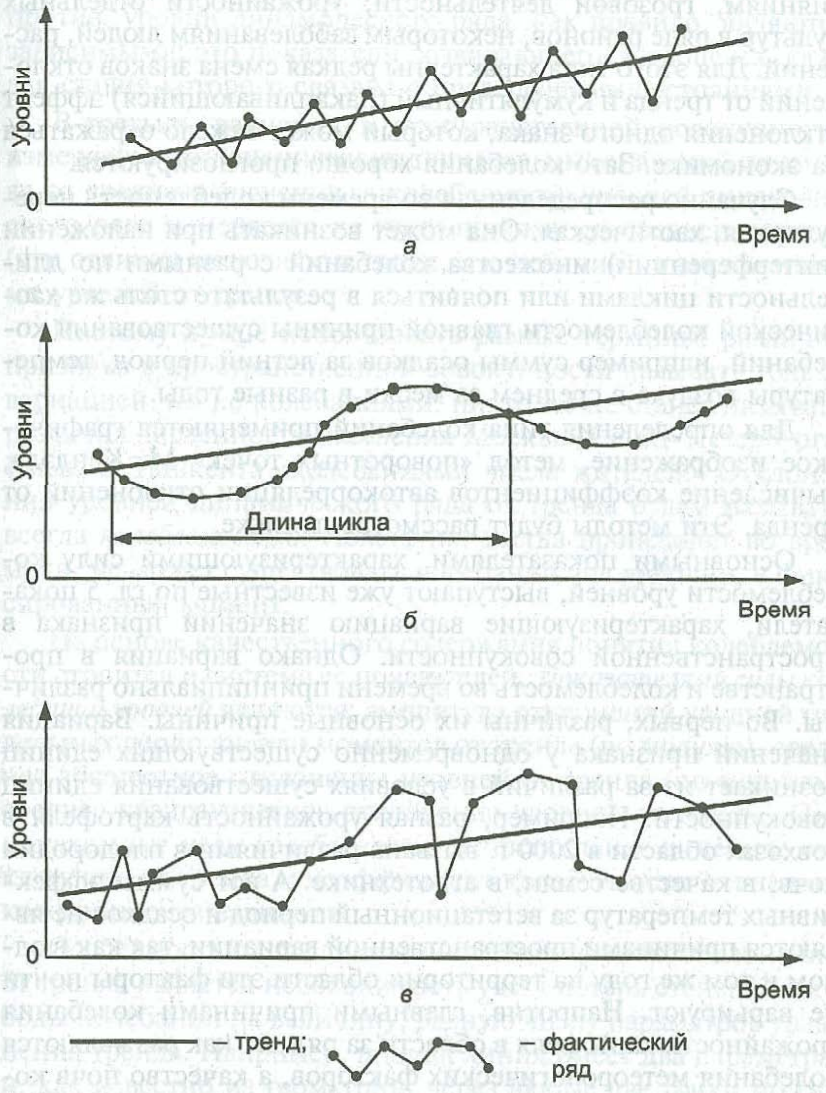


Рис. 12.2. Виды колебаний:
а — пилообразная, или маятниковая; *б* — долгопериодическая циклическая; *в* — случайно распределенная во времени

сияниям, грозовой деятельности, урожайности отдельных культур в ряде районов, некоторым заболеваниям людей, растений. Для этого типа характерны редкая смена знаков отклонений от тренда и кумулятивный (накапливающийся) эффект отклонения одного знака, который может тяжело отражаться на экономике. Зато колебания хорошо прогнозируются.

Случайно распределенная во времени колеблемость нерегулярная, хаотическая. Она может возникать при наложении (интерференции) множества колебаний с разными по длительности циклами или появиться в результате столь же хаотической колеблемости главной причины существования колебаний, например суммы осадков за летний период, температуры воздуха в среднем за месяц в разные годы.

Для определения типа колебаний применяются графическое изображение, метод «поворотных точек» М. Кендэла, вычисление коэффициентов автокорреляции отклонений от тренда. Эти методы будут рассмотрены ниже.

Основными показателями, характеризующими силу колеблемости уровней, выступают уже известные по гл. 5 показатели, характеризующие вариацию значений признака в пространственной совокупности. Однако вариация в пространстве и колеблемость во времени принципиально различны. Во-первых, различны их основные причины. Вариация значений признака у одновременно существующих единиц возникает из-за различий в условиях существования единиц совокупности. Например, разная урожайность картофеля в совхозах области в 2000 г. вызвана различиями в плодородии почв, в качестве семян, в агротехнике. А вот суммы эффективных температур за вегетационный период и осадков не являются причинами пространственной вариации, так как в одном и том же году на территории области эти факторы почти не варьируют. Напротив, главными причинами колебания урожайности картофеля в области за ряд лет как раз являются колебания метеорологических факторов, а качество почв колебаний почти не имеет. Что же касается общего прогресса агротехники, то он является причиной тренда, но не колеблемости.

Во-вторых, коренное отличие состоит в том, что значения варьирующего признака в пространственной совокупности можно считать в основном не зависимыми друг от друга, на-

против, уровни динамического ряда, как правило, являются зависимыми: это показатели развивающегося процесса, каждая стадия которого связана с предыдущими состояниями.

В-третьих, вариация в пространственной совокупности измеряется отклонениями индивидуальных значений признака от среднего значения, а колеблемость уровней динамического ряда измеряется не их отличиями от среднего уровня (эти отличия включают и тренд, и колебания), а отклонениями уровней от тренда.

Поэтому лучше использовать разные термины: различия признака в пространственной совокупности называть только вариацией, но не колебаниями: никто же не станет называть различия численности населения Москвы, Санкт-Петербурга, Киева и Ташкента «колебаниями числа жителей!» Отклонения уровней динамического ряда от тренда будем называть всегда *колеблемостью*. Колебания всегда происходят во времени, не может существовать колебаний вне времени, в фиксированный момент.

На основе качественного содержания понятия колеблемости строится и система ее показателей. *Показателями силы колебаний уровней* являются: амплитуда отклонений уровней отдельных периодов или моментов от тренда (по модулю), среднее абсолютное отклонение уровней от тренда (по модулю), среднее квадратическое отклонение уровней от тренда. *Относительные меры колеблемости*: относительное линейное отклонение от тренда и коэффициент колеблемости — аналоги коэффициента вариации.

Особенностью методики вычисления средних отклонений от тренда является необходимость учета потерь степеней свободы колебаний на величину, равную числу параметров уравнения тренда. Например, прямая линия имеет два параметра, и, как известно из геометрии, через любые две точки можно провести прямую линию. Значит, имея лишь два уровня, мы проведем линию тренда точно через эти два уровня, и никаких отклонений уровней от тренда не окажется, хотя на самом деле и эти два уровня включали колебания, не были свободны от действия факторов колеблемости. Парабола 2-го порядка пройдет точно через любые три точки и т.п.

Учитывая потерю степеней свободы, основные абсолютные показатели колеблемости вычисляются по формулам (12.33) и (12.34):

среднее линейное отклонение $$a(t) = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n-p}; \quad (12.33)$$

среднее квадратическое отклонение $$s(t) = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-p}}, \quad (12.34)$$

где n — число уровней;

y_i — фактический уровень;

$\hat{y}_i$ — выравненный уровень, тренд;

p — число параметров тренда.

Знак времени «t» в скобках после обозначения показателя означает, что это показатель не обычной пространственной вариации, как в гл. 5, а показатель колеблемости во времени.

Относительные показатели колеблемости вычисляются делением абсолютных показателей на средний уровень за весь изучаемый период. Расчет показателей колеблемости проведем по результатам анализа динамики индекса цен (см. табл. 12.8). Тренд примем по результатам многократного скользящего выравнивания, т.е. $\hat{y} = 104{,}3 - 1{,}433t$; $t = 0$ в 1987 г.

1. Амплитуда колебаний составила от −14,0 в 1986 г. до +15,2 в 1984 г., т.е. 29,2 пункта.

2. Среднее линейное отклонение по модулю найдем, сложив модули $|u_i|$ (их сумма равна 132,3) и разделив на $(n - p)$, согласно формуле (12.33):

$$a(t) = \frac{132{,}2}{17-2} = 8{,}82 \text{ пункта.}$$

3. Среднее квадратическое отклонение уровней от тренда по формуле (12.34) составило:

$$s(t) = \sqrt{\frac{1340{,}41}{17-2}} = 9{,}45 \text{ пункта.}$$

Небольшое превышение среднего квадратического отклонения над линейным указывает на отсутствие среди отклонений, резко выделяющихся по абсолютной величине.

4. Коэффициент колеблемости: $v_{(t)} = \frac{s(t)}{\bar{y}} = \frac{9{,}45}{104{,}53} = 0{,}0904$, или 9,04%. Колеблемость умеренная, несильная. Для сравнения приведем показатели (без расчетов) по колебаниям урожайности картофеля (данные табл. 12.1 и 12.5) — среднее квадратическое отклонение от линейного тренда: $s(t) = 14{,}38$ ц с 1 га, коэффициент колеблемости: $v(t) = 8{,}35\%$.

Для выявления типа колебаний воспользуемся приемом, предложенным М. Кендэлом. Он состоит в подсчете так называемых поворотных точек в ряду отклонений от тренда u_i, т.е. локальных экстремумов. Отклонение, либо большее по алгебраической величине, либо меньшее двух соседних, отмечается точкой. Обратимся к рис. 12.2. При маятниковой колеблемости все отклонения, кроме двух крайних, будут поворотными, следовательно, их число составит $n - 2$. При долгопериодических циклах на цикл приходятся один минимум и один максимум, а *общее число точек* составит: $2(n : l)$, где l — длительность цикла. При случайно распределенной во времени колеблемости, как доказал М. Кендэл, число поворотных точек в среднем составит: $2/3\ (n - 2)$. В нашем примере при маятниковой колеблемости было бы 15 точек, при связанной с 11-летним циклом колеблемости было бы $2 \cdot (17 : 11) \approx 3$ точки, при случайно распределенной во времени в среднем было бы $2/3(17 - 2) = 10$ точек. Фактическое число точек — 6 выходит за границы двукратного среднего квадратического отклонения числа поворотных точек, которое, по Кендэлу, равно: $\sqrt{\frac{16n-29}{90}}$[1].

В нашем примере: $\sqrt{\frac{16 \cdot 17 - 29}{90}} = 1{,}6$.

Наличие 6 точек при 2 точках за цикл означает, что в ряду могут быть примерно 3 цикла, продолжительность периода которых 5,5—6 лет. Возможно сочетание таких циклических колебаний со случайными.

[1] *Кендэл М., Юл Дж. Э.* Теория статистики. — М.: Госстатиздат, 1960. — С. 708.

Другой метод анализа типа колеблемости и поиска длины цикла основан на вычислении *коэффициентов автокорреляции отклонений от тренда*.

Автокорреляция — это корреляция между уровнями ряда или отклонениями от тренда, взятыми со сдвигом во времени: на 1-й период (год), на 2-й, на 3-й и т.д., поэтому говорят о коэффициентах автокорреляции разных *порядков*: первого, второго и т.д. Рассмотрим сначала коэффициент автокорреляции отклонений от тренда первого порядка.

Одна из основных формул для расчета коэффициента автокорреляции отклонений от тренда имеет вид:

$$r_u^{a_1} = \frac{\sum_{i=1}^{n-1} u_i u_{i+1}}{\frac{u_1^2}{2} + \sum_{i=2}^{n-1} u_i^2 + \frac{u_n^2}{2}} \quad (12.35)$$

Как легко видеть по табл. 12.7, первое и последнее в ряду отклонения участвуют только в одном произведении в числителе, а все прочие отклонения от второго до $(n-1)$-го — в двух. Поэтому и в знаменателе квадраты первого и последнего отклонений следует взять с половинным весом, как в хронологической средней. По данным табл. 12.8 имеем:

$$r_u^{a_1} = \frac{446{,}2}{60{,}5 + 1051{,}1 + 83{,}2} = 0{,}373.$$

Теперь обратимся к рис. 12.2. При маятниковой колеблемости все произведения в числителе будут отрицательными величинами и коэффициент автокорреляции первого порядка будет близок к -1. При долгопериодических циклах будут преобладать положительные произведения соседних отклонений, а смена знака происходит лишь дважды за цикл. Чем длиннее цикл, тем больше перевес положительных произведений в числителе и коэффициент автокорреляции первого порядка ближе к $+1$. При случайно распределенной во времени колеблемости знаки отклонений чередуются хаотически, число положительных произведений близко к числу отрицательных, ввиду чего коэффициент автокорреляции близок к нулю. Полученное значение говорит о наличии как случайно распределенных во времени колебаний, так и циклических. Коэффициент автокорреляции следующих порядков: II = $-0{,}577$; III = $-0{,}611$; IV = $-0{,}095$; V = $+0{,}376$; VI = $+0{,}404$; VII = $+0{,}044$. Следовательно, противофаза цикла ближе всего к 3 годам (наибольший отрицательный коэффициент при сдвиге на 3 года), а совпадающие фазы ближе к 6 годам, что и дает длину цикла колебаний. Максимальные по абсолютной величине коэффициенты неблизки к единице. Это означает, что циклическая колеблемость смешана со значительной случайной колеблемостью. Таким образом, подробный автокорреляционный анализ в целом дал те же результаты, что и выводы по автокорреляции первого порядка.

Если динамический ряд достаточно длинен, можно поставить и решить задачу об изменении показателей колеблемости с течением времени. Для этого рассчитывают эти показатели по подпериодам, но длительностью не менее 9–11 лет, иначе измерения колеблемости ненадежны. Кроме того, можно рассчитывать показатели колеблемости скользящим способом, а затем провести их выравнивание, т.е. вычислить тренд показателей колеблемости. Это полезно для вывода о действенности мер, применявшихся для уменьшения колебаний урожайности и других нежелательных колебаний, а также для того, чтобы по тренду сделать прогноз ожидаемых в будущем размеров колебаний.

12.9. Измерение устойчивости в динамике

Понятие «устойчивость» используется в различных смыслах. По отношению к статистическому изучению динамики рассмотрим два аспекта этого понятия: 1) устойчивость как категория, противоположная колеблемости; 2) устойчивость направленности изменений, т.е. устойчивость тенденции.

В первом понимании *показатель устойчивости*, который может быть только относительным, должен изменяться от нуля до единицы (100%). Это разность между единицей и относительным показателем колеблемости: $1 - v(t)$. Если коэффициент колеблемости составил 9,0%, то коэффициент устойчивости равен 100% – 9,0% = 91,0%. Этот показатель характеризует близость фактических уровней к тренду и совершенно не зависит от характера последнего. Слабая колеблемость

и высокая устойчивость уровней в данном смысле могут существовать даже при полном застое в развитии, когда тренд выражен горизонтальной прямой.

Устойчивость во втором смысле характеризует не сами по себе уровни, а процесс их направленного изменения. Можно узнать, например, насколько устойчив процесс сокращения удельных затрат ресурсов на производство единицы продукции, является ли устойчивой тенденция снижения детской смертности и т.д. С этой точки зрения полной устойчивостью направленного изменения уровней динамического ряда следует считать такое изменение, в процессе которого каждый следующий уровень либо выше всех предшествующих (устойчивый рост), либо ниже всех предшествующих (устойчивое снижение). Всякое нарушение строго ранжированной последовательности уровней свидетельствует о неполной устойчивости изменений.

Из определения понятия устойчивости тенденции вытекает и метод определения ее показателя. В качестве показателя устойчивости можно использовать *коэффициент корреляции рангов* Ч. Спирмена ρ (см. гл. 10):

$$\rho = 1 - \frac{6\sum_{i=1}^{n}\Delta_i^2}{n^3 - n}, \qquad (12.36)$$

где n — *число уровней*;

Δ_i — разность рангов уровней и номеров периодов времени.

При полном совпадении рангов уровней, начиная с наименьшего, и номеров периодов (моментов) времени по их хронологическому порядку коэффициент корреляции рангов равен +1. Это значение соответствует случаю полной устойчивости возрастания уровней. При полной противоположности рангов уровней рангам лет коэффициент Спирмена равен −1, что означает полную устойчивость процесса сокращения уровней. При хаотическом чередовании рангов уровней коэффициент близок к нулю, это означает неустойчивость какой-либо тенденции. Приведем расчет коэффициента корреляции Спирмена по данным о динамике индекса цен (табл. 12.5) в табл. 12.9.

Таблица 12.9

Расчет коэффициентов корреляции рангов Спирмена

Год	Уровни y_i	Ранг лет P_x	Ранг уровней P_y	$P_x - P_y$	$(P_x - P_y)^2$
1979	105	1	8	7	49
1980	111	2	13	11	121
1981	110	3	12	9	81
1982	106	4	9,5	5,5	30,25
1983	118	5	16	11	121
1984	124	6	17	11	121
1985	113	7	14,5	7,5	56,25
1986	92	8	3,5	4,5	20,25
1987	91	9	1,5	7,5	56,25
1988	109	10	11	1	1
1989	113	11	14,5	3,5	12,25
1990	100	12	6	6	36
1991	94	13	5	8	64
1992	91	14	1,5	12,5	156,25
1993	92	15	3,5	11,5	132,25
1994	102	16	7	9	81
1995	106	17	9,5	7,5	56,25
Σ	1777	—	—	—	1141

Ввиду наличия трех пар «связанных рангов» применяем формулу (11.26):

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(R_x - R_y)^2 - A}{\sqrt{(n^3 - n)(n^3 - n - A)}} = 1 - \frac{6 \cdot 1141 - 2}{\sqrt{(17^3 - 17)(17^3 - 17 - 2)}} = -0{,}398.$$

Отрицательное значение ρ указывает на наличие тенденции снижения уровней, причем устойчивость этой тенденции ниже средней.

При этом следует иметь в виду, что даже при 100%-ной устойчивости тенденции в ряду динамики может быть колеблемость уровней и коэффициент их устойчивости будет ниже 100%. При слабой колеблемости, но еще более слабой тенденции, напротив, возможен высокий коэффициент устойчивости уровней но близкий к нулю коэффициент устойчивости

тренда. В целом же оба показателя связаны, конечно, прямой зависимостью: чаще всего большая устойчивость уровней наблюдается одновременно с большей устойчивостью тренда.

Устойчивость тенденции развития, или комплексная устойчивость в динамике, может быть охарактеризована соотношением между среднегодовым абсолютным изменением и средним квадратическим (либо линейным) отклонением уровней от тренда:

$$c = \frac{b}{s(t)}. \qquad (12.37)$$

Если, как нередко бывает, распределение отклонений уровней ряда от тренда близко к нормальному, то с вероятностью 0,95 отклонение от тренда вниз не превысит 1,645 $s(t)$ по величине. Следовательно, если в ряду динамики $c > 1{,}64$, то уровни, более низкие, чем предыдущие, в среднем будут встречаться менее 5 раз за 100 периодов, или 1 раз из 20, т.е. устойчивость тренда будет высока. При $c = 1$ нарушения ранжированности уровней будут встречаться в среднем 16 раз из 100, а при $c = 0{,}5$ — уже 31 раз из 100, т.е. устойчивость тенденции будет низкой. Можно также пользоваться отношением среднего темпа прироста к коэффициенту колеблемости, что дает показатель, близкий к c-показателю устойчивости. Этот показатель более пригоден для экспоненциального тренда. О показателях устойчивости нелинейных трендов и об общих проблемах устойчивости экономических и социальных процессов можно подробнее прочесть в рекомендуемой к данной главе литературе [2].

12.10. Сезонные колебания и полное разложение дисперсии уровней динамического ряда

Сезонными называют периодические колебания, возникающие под влиянием смены времени года. Их роль очень велика в агропромышленном комплексе, торговле, заболеваемости, строительстве, деятельности рекреационных учреждений, на транспорте. Сезонные колебания строго цикличны — повторяются через каждый год, хотя сама длительность времен года имеет колебания. Для изучения сезонных колебаний

необходимо иметь уровни за каждый квартал, а лучше за каждый месяц, иногда даже за декады, хотя декадные уровни могут уже сильно исказиться мелкомасштабной случайной колеблемостью.

Следует еще раз указать, что не всякие различия в месячных или квартальных уровнях являются сезонными колебаниями, а только регулярно повторяющиеся. Если же различия месячных уровней или любых внутригодичных уровней в один год распределены совершенно иначе, чем в другой год, то это не сезонные, а случайные колебания, т.е. колебания, вызванные причинами, не связанными со сменой времен года. Например, такими могут быть колебания курсов акций, обменных курсов валют, вызванные изменением финансовой политики государства, научно-техническими открытиями, политическими кризисами в стране и мире, слиянием и разделением компаний и т.п.

Поскольку интервальные уровни зависят от длительности интервалов времени, а длина месяцев не равная, правильнее анализировать колебания не по фактическим месячным уровням, а по уровням, пересчитанным на равную (30-дневную) длительность всех месяцев, или среднесуточным. Если изучаются сезонные колебания за отдельный год, то обычно тренд не принимается во внимание, и отклонения месячных (30-дневных) уровней исчисляются от среднемесячного уровня за год. Кроме рассмотренных показателей колеблемости для характеристики сезонных колебаний важное значение имеет форма сезонной «волны», изучаемая с помощью относительных показателей — отношений месячных уровней к среднемесячному (так называемый индекс сезонности). Лучше, конечно, изучать сезонные колебания за несколько лет, чтобы сгладить случайные колебания и точнее измерить сезонные.

Если сезонная колеблемость имеет синусоидальный характер, т.е. плавно изменяется в течение года от минимума до максимума и обратно, для ее моделирования пригодна тригонометрическая модель вида

$$\hat{Y} = a + b_1 \sin \alpha + \cos \alpha, \qquad (12.38)$$

где α — угол, получаемый для каждого месяца нарастающим итогом.

Таблица 12.10

Расчет тригонометрической модели сезонности продуктивности коров

Месяц	i_j, %	α °	sin α	cos α	$i_j$sin α	$i_j$cos α	$\hat{i}_j$, %	$i_j - \hat{i}_j$	$(i_j - \hat{i}_j)^2$	$i_j - 100$	$(i_j - 100)^2$
А	1	2	3	4	5	6	7	8	9	10	11
Январь	36	0	0	1	0	36	38,2	−2,2	4,84	−61,8	3819
Февраль	60	30	0,5	0,866	30,0	52	50,9	9,1	82,81	−49,1	2411
Март	81	60	0,866	0,5	70,1	40,5	76,7	4,3	18,49	−23,3	543
Апрель	100	90	1	0	100,0	0	108,8	−8,8	77,44	8,8	77
Май	129	120	0,866	−0,5	111,7	−64,5	138,5	−9,5	90,25	38,5	1482
Июнь	163	150	0,5	−0,866	81,5	−141,2	157,9	5,1	26,01	57,9	3352
Июль	175	180	0	−1	0	−175,0	161,8	13,2	174,24	61,8	3819
Август	148	210	−0,5	−0,866	−74	−128,2	149,1	1,1	1,21	49,1	2411
Сентябрь	120	240	−0,866	−0,5	−103,9	−60	123,3	−3,3	10,89	23,3	543
Октябрь	90	270	−1	0	−90	0	91,2	−1,2	1,44	−8,8	77
Ноябрь	56	300	−0,866	0,5	−48,5	28	61,5	−5,5	30,25	−38,5	1482
Декабрь	48	330	−0,5	0,866	−24,0	41,6	42,1	5,9	34,81	−57,9	3352
Σ	1200	X	0	0	52,9	−370,8	1200	0	552,68	0	23 368

Примечание. $\alpha = i_j$ за год = 100%.

«Круг года» считается равным 360 °, а каждый месяц равен: 360 ° : 12 = 30 °. Таким образом, для января $\alpha = 0$ °, для февраля $\alpha = 30$ °, для марта $\alpha = 60$ ° и т.д. Такой характер динамики имеет, например, молочная продуктивность коров, повышающаяся от зимы к лету, а затем — падающая; реализация товаров летнего (или зимнего) ассортимента; численность отдыхающих на южных курортах и т.д. Алгоритм построения модели сезонных колебаний таков.

1. За каждый год вычисляются отношения месячных уровней к среднемесячному $i_j = y_i : \bar{y}$.

2. Для получения типичной картины сезонных колебаний эти отношения для каждого месяца усредняются за ряд лет (не менее 5—8).

Тем самым исключаются и тренд, и большая часть случайных особенностей в отдельные годы. Расчет представлен в табл. 12.10.

$$b_1 = \frac{\sum_{j}^{12} \bar{i}_j \cdot \sin\alpha}{6} = \frac{52,9}{6} = 8,82;$$

$$b_2 = \frac{\sum_{j}^{12} \bar{i}_j \cdot \cos\alpha}{6} = \frac{-370,8}{6} = -61,8;$$

$$\hat{y} = 100 + 8,82 \sin\alpha - 61,8 \cos\alpha.$$

Как видно по данным графы 8, модель довольно хорошо отражает фактическую сезонность. Это же видно и на рис. 12.3. Расхождение величины фактических индексов сезонности и рассчитанных по модели можно трактовать как остаточную колеблемость, не отражаемую моделью. Она измеряется суммой квадратов отклонений $\Sigma(i_j - \hat{i})^2$. Для данного примера остаточная колеблемость равна 552,68. Сезонная колеблемость измеряется суммой квадратов отклонений расчетных значений индексов сезонности i_j от средней, т.е. от 100%. Эта сумма в итоге графы 11 равна 23 368. Число степеней свободы вариации для модели, имеющей две независимые переменные, равно двум, для остаточных: 12 − 1 − 2 = 9. Составим таблицу дисперсионного анализа (табл. 12.11).

Таблица 12.11

Дисперсионный анализ модели сезонности

Источник вариации	Сумма квадратов D	Число степеней свободы $d.f.$	Дисперсия $s^2 = D/d.f.$	F-критерий
Модель	23368	2	11684	190,3
Остаточная	553	9	61,4	1
Всего	23921	11	—	

Табличное значение критерия Фишера для двух и девяти степеней свободы при уровне значимости 0,01 равно 8,02. Таким образом, надежность модели очень высокая, наличие сезонности доказано. Не следует, впрочем, забывать, что большая часть случайной колеблемости была погашена на втором шаге расчетов — при усреднении индексов сезонности за несколько лет. Но даже если было погашено и 80% случайной колеблемости (т.е. F-критерий был бы в 5 раз меньше), вывод остается неизменным.

Наиболее полную и точную методику анализа сезонных колебаний рассмотрим на примере производства молока в России за IV кв. 1998 г. — I кв. 2002 г. (табл. 12.12).

Как видно из табл. 12.12, сезонные колебания производства молока состоят в его повышении во II и III кварталах и понижении в IV и I кварталах. Вычислим параметры уравнения тренда за период IV кв. 1998 г. — IV кв. 2001 г. Получим: $\hat{Y} = 7{,}905 + 0{,}095 \cdot t$, где $t = 0$ во II кв. 2002 г., т.е. средний квартальный прирост составит 95 тыс. т за квартал. Если вычислить тренд для периода I кв. 1999 г. — I кв. 2002 г., то получим уравнение: $\hat{Y} = 7{,}976 - 0{,}062 \cdot t$, где $t = 0$ в III кв. 2000 г., т.е. средний прирост объема производства отрицателен (в среднем −62 тыс. т за квартал). В первом случае был получен завышенный среднеквартальный прирост, потому что в начале базы расчета тренда находятся два низких уровня, в конце — один, а высокие уровни сдвинуты по всей базе ближе к концу ряда, нежели к началу базы. Во втором случае в начале базы расчета тренда находится один низкий уровень, а в конце — два подряд, высокие же уровни вторых и третьих кварталов расположены ближе к началу базы, отчего средний прирост занижен.

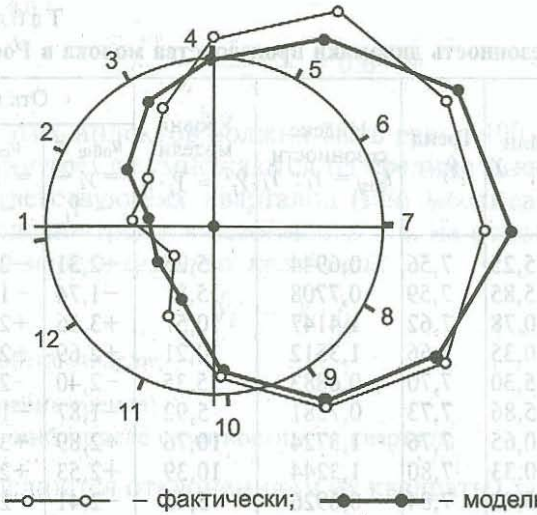


Рис. 12.3. Сезонные колебания продуктивности коров: 1—12 — месяцы года

Для того чтобы при наличии существенных сезонных колебаний избежать искажения параметров тренда, следует так расположить начало и конец базы расчета, чтобы сезонные подъемы и снижение уровней располагались симметрично по отношению к началу и к концу базы расчета тренда. В данном ряду следует вычислить тренд за период от IV кв. 1998 г. до I кв. 2002 г. Тогда от начала до конца базы равномерно расположатся два низких уровня, затем два высоких, опять два низких и т.д. до двух низких в конце. Уравнение тренда в этом случае таково: $\hat{Y} = 7{,}785 + 0{,}035 \cdot t$, где $t = +0{,}5$ в III кв. 2000 г.

Анализ сезонности проводится по данным табл. 12.12 (и аналогичным ей) по следующей методике.

1. Вычисляется уравнение тренда, учитывая сказанное выше о выборе базы для расчета; рассчитываются по уравнению выравненные уровни, $\hat{y}_i$.

2. Каждый фактический уровень квартального производства делится на уровень тренда (выравненный) для того же периода, получаются индивидуальные индексы сезонности: $i_{сезjk} = Y_i : \hat{Y}_i$, т.е. показатели сезонности для j-го квартала в k-м году.

Таблица 12.12

Сезонность динамики производства молока в России

Год, квартал	Млн тонн Y_i	Тренд $\hat{Y}_i$	Индекс сезонности $i_{сезj} = Y_i : \hat{Y}_i$	Уровни модели $\hat{Y}_i' = \bar{Y}_i \cdot \bar{i}_j$	Отклонения		
					$u_{общ} = y_i - \hat{Y}_i$	$u_{сез} = \hat{Y}_i' - \hat{Y}_i$	$u_{случ} = \hat{Y}_i - \hat{Y}_i'$
1998, IV	5,25	7,56	0,6944	5,25	−2,31	−2,31	0
1999, I	5,85	7,59	0,7708	5,81	−1,74	−1,78	+0,04
II	10,78	7,62	1,4147	10,57	+3,16	+2,95	+0,21
III	10,35	7,66	1,3512	10,21	+2,69	+2,55	+0,14
IV	5,30	7,70	0,6883	5,35	−2,40	−2,35	−0,05
2000, I	5,86	7,73	0,7581	5,92	−1,87	−1,81	−0,06
II	10,65	7,76	1,3724	10,76	+2,89	+3,00	−0,11
III	10,33	7,80	1,3244	10,39	+2,53	+2,59	−0,06
IV	5,43	7,84	0,6926	5,45	−2,41	−2,39	−0,02
2001, I	5,94	7,87	0,7548	6,02	−1,93	−1,85	−0,08
II	10,86	7,90	1,3747	10,96	+2,96	+3,06	−0,10
III	10,49	7,94	1,3212	10,58	+2,55	+2,64	−0,09
IV	5,62	7,98	0,7043	5,55	−2,36	−2,43	+0,07
2002, I	6,23	8,01	0,7778	6,13	−1,78	−1,88	+0,10
Итого	108,99	108,96		108,95	0,03	0,01	0,04

Источник. Вопросы статистики. — 2002. — № 9. — С. 43.

3. Индивидуальные индексы сезонности, включающие и случайные отклонения, осредняются по кварталам, т.е. рассчитываются *средние индексы сезонности* $\bar{i}_{сезj}$ для каждого j-го квартала (независимо от особенностей года).

Для I кв.:

$$i_I = \frac{\sum i_{сезI}}{4} = \frac{0{,}7708 + 0{,}7581 + 0{,}7548 + 0{,}7778}{4} = 0{,}7654.$$

Для II кв.:

$$i_{II} = \frac{\sum i_{сезII}}{3} = 1{,}3873.$$

Для III кв.:

$$i_{III} = \frac{\sum i_{сезIII}}{3} = 1{,}3323.$$

Для IV кв.:

$$i_{IV} = \frac{\sum i_{сезIV}}{3} = 0{,}6949.$$

Сумма этих индексов должна быть равна 400.

4. Уровни тренда умножаются на средние индексы сезонности соответствующих кварталов (или месяцев), получаем уровень модели=тренд · сезонность, т.е. на основе *мультипликативной модели сезонных колебаний*:

$$\hat{Y}_i' = \hat{Y}_i \cdot \bar{i}_j,$$

где $\hat{Y}_i'$ — уровень модели;

$\hat{Y}_i$ — уровень тренда;

i_j — средний индекс сезонности j-го квартала.

5. Вычисляются отклонения (и их квадраты) за счет сезонности:

$$u_{сезij} = \hat{Y}_{ij}' - \hat{Y}_{ij}.$$

6. Вычисляются отклонения (и их квадраты) за счет случайной колеблемости:

$$u_{случij} = Y_{ij} - \hat{Y}_{ij}'.$$

7. Вычисляются общие отклонения:

$$u_{общ} = u_{сез} + u_{случ} = Y_{ij} - \hat{Y}_{ij}.$$

Итоги по графам $u_{общ}$, $u_{сез}$, $u_{случ}$ в принципе должны быть нулевыми. Несущественные расхождения, как в нашем примере, из-за округлений при расчетах не имеют значения.

Графическое изображение временно́го ряда с наличием сезонных колебаний в системе прямоугольных координат возможно двумя способами. Первый способ показан на рис. 12.4. На рис. 12.4 видно характерное для сезонных колебаний регулярное чередование повышения и понижения уровней ряда. Второй способ графического изображения сезонности — в полярных координатах: средний уровень принимается за длину радиуса-вектора ρ, а год — за окружность. Этот способ хорош прежде всего при наличии месячных данных. Каждому месяцу соответствует 30 °. Концы отрезков, соответствующих

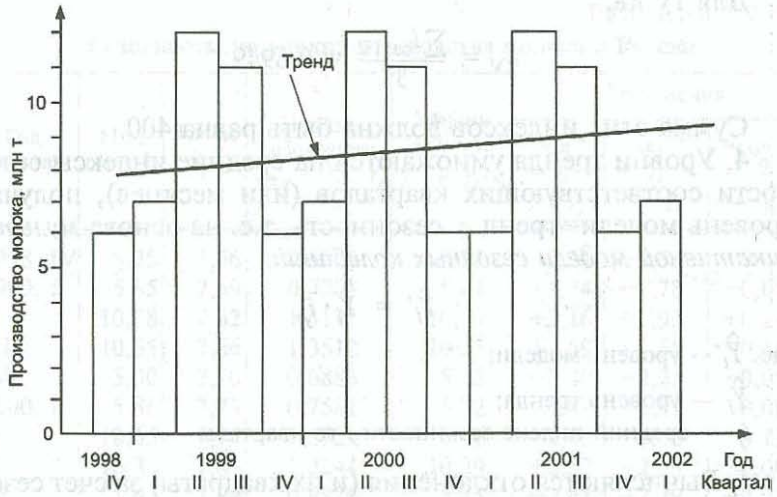


Рис. 12.4. Динамика производства молока в России

по длине месячным индексам сезонности и откладываемых от центра окружности (круга), соединяют ломаной линией, которая в период сезонных повышений выходит за окружность, а в период сезонных понижений «втягивается» внутрь окружности (рис. 12.3). Такой график называется *радиальной диаграммой*. По квартальным данным подобный график невыразителен (рис. 12.5, *а*). Сезонность может быть представлена в виде графика сезонной волны, на котором изображаются индексы сезонности (рис. 12.5, *б*).

Кроме мультипликативной модели сезонных колебаний может быть построена *аддитивная модель*, т.е. такая модель, в которой сезонные повышение и понижение уровней выражаются слагаемыми, соответственно положительными или отрицательными, добавляемыми к уровням тренда. Для того чтобы реализовать такую модель на ПЭВМ, каждый квартал (или месяц) обозначается особый «структурной» (иногда ее называют «фиктивной») переменной, принимающей для данного квартала или месяца значение «1», а для всех других значение «0». Сущность метода структурных переменных излагается в главе о корреляции и регрессии (гл. 9).

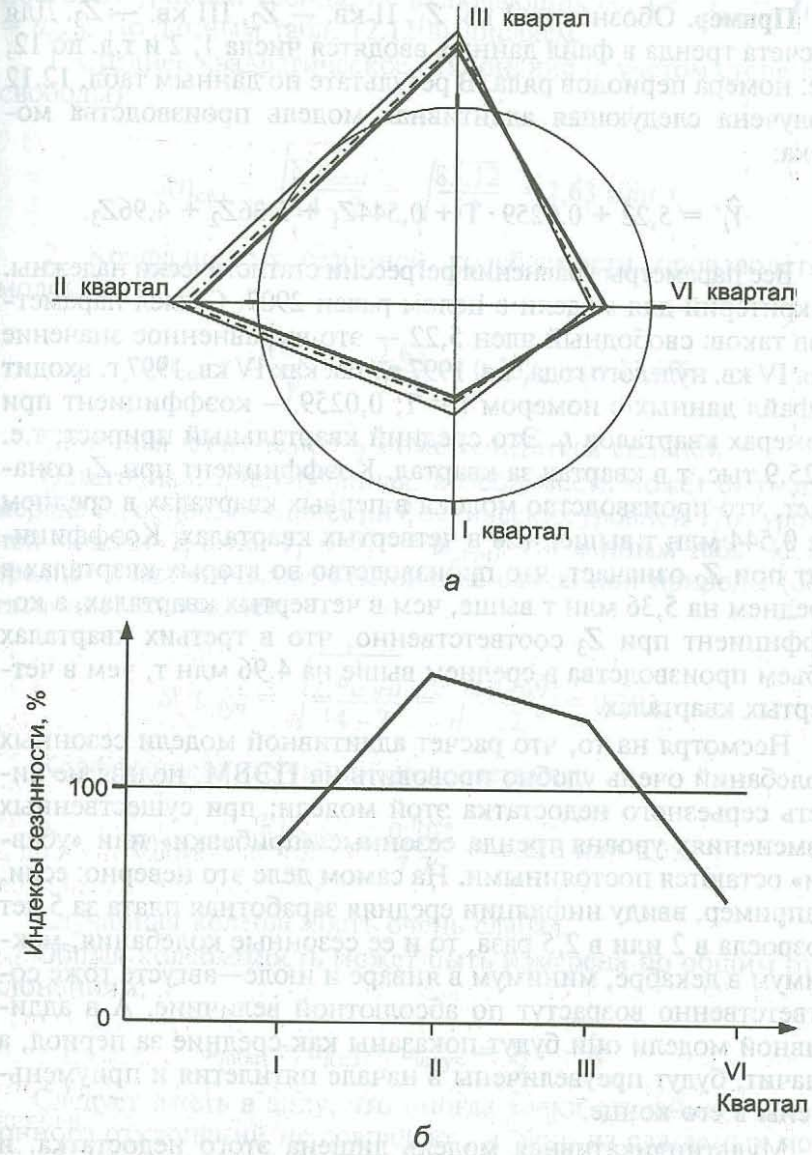


Рис. 12.5. Сезонность производства молока в России:
а — радиальная диаграмма; *б* — график сезонной волны

Пример. Обозначим I кв. Z_1, II кв. — Z_2, III кв. — Z_3. Для расчета тренда в файл данных вводятся числа 1, 2 и т.д. до 12, т.е. номера периодов ряда. В результате по данным табл. 12.12 получена следующая аддитивная модель производства молока:

$$\hat{Y}_i' = 5{,}22 + 0{,}0259 \cdot \mathrm{T} + 0{,}544Z_1 + 5{,}36Z_2 + 4{,}96Z_3.$$

Все параметры уравнения регрессии статистически надежны. F-критерий для модели в целом равен 2907. Смысл параметров таков: свободный член 5,22 — это выравненное значение для IV кв. нулевого года, т.е. 1997 г., так как IV кв. 1997 г. входит в файл данных с номером $t = 1$; 0,0259 — коэффициент при номерах кварталов t_i. Это средний квартальный прирост, т.е. +25,9 тыс. т в квартал за квартал. Коэффициент при Z_1 означает, что производство молока в первых кварталах в среднем на 0,544 млн т выше, чем в четвертых кварталах. Коэффициент при Z_2 означает, что производство во вторых кварталах в среднем на 5,36 млн т выше, чем в четвертых кварталах, а коэффициент при Z_3 соответственно, что в третьих кварталах объем производства в среднем выше на 4,96 млн т, чем в четвертых кварталах.

Несмотря на то, что расчет аддитивной модели сезонных колебаний очень удобно проводить на ПЭВМ, нельзя не видеть серьезного недостатка этой модели: при существенных изменениях уровня тренда сезонные «прибавки» или «убавки» остаются постоянными. На самом деле это неверно: если, например, ввиду инфляции средняя заработная плата за 5 лет возросла в 2 или в 2,5 раза, то и ее сезонные колебания, максимум в декабре, минимум в январе и июле—августе тоже соответственно возрастут по абсолютной величине. А в аддитивной модели они будут показаны как средние за период, а значит, будут преувеличены в начале пятилетия и приуменьшены в его конце.

Мультипликативная модель лишена этого недостатка, и при существенном тренде она гораздо точнее отразит сезонные колебания, чем аддитивная модель.

Сила и интенсивность сезонных колебаний измеряются на основе отклонений уровней модели, включающих сезонность, от уровней тренда, не включающих ее, т.е. $\hat{Y}_i' - \hat{Y}_i = u_{\text{сез}i}$. По данным табл. 12.12 вычисляем.

1. Среднее квадратическое отклонение (с учетом степеней свободы):

$$S(t)_{\text{сез}} = \sqrt{\frac{\sum u^2_{\text{сез}i}}{14-2}} = \sqrt{\frac{83{,}12}{12}} = 2{,}63 \text{ млн т.}$$

2. Коэффициент сезонной колеблемости производства молока:

$$v(t)_{\text{сез}} = \frac{S(t)_{\text{сез}}}{\bar{Y}} = \frac{2{,}63}{7{,}78} = 0{,}338, \text{ или } 33{,}8\%.$$

Сезонная колеблемость может считаться сильной.

Остаточная, или случайная, колеблемость может быть измерена на основе отклонений фактических уровней Y_i от уровней модели $\hat{Y}_i'$, т.е. $Y_i - \hat{Y}_i' = u_{\text{случ}i}$. По данным табл. 12.12 среднее квадратическое отклонение случайной природы (остаточное) составляет:

$$S(t)_{\text{случ}} = \sqrt{\frac{\sum u^2_{\text{случ}i}}{14-2}} = \sqrt{\frac{0{,}1269}{12}} = 0{,}103.$$

Коэффициент случайной колеблемости:

$$v(t)_{\text{случ}} = \frac{S(t)_{\text{случ}}}{\bar{y}} = \frac{0{,}103}{7{,}78} = 0{,}013 \text{ или } 1{,}3\%.$$

Случайная колеблемость очень слабая.

Общая колеблемость может быть измерена по общим отклонениям:

$$u_{i\text{общ}} = u_{i\text{сез}} + u_{i\text{случ}} = Y_i - \hat{Y}_i.$$

Следует иметь в виду, что иногда знаки случайного и сезонного отклонений не совпадают, и одно из слагаемых может по абсолютной величине быть больше суммы, а уж тем более его квадрат. Если бы между случайной и сезонной колеблемостью была обратная зависимость (а на выбранном ограниченном отрезке времени это может случиться), то общая

колеблемость может оказаться меньше сезонной. При преобладании прямой связи между сезонными и случайными отклонениями, наоборот, общая колеблемость будет больше той, которая была бы при отсутствии связи между сезонными и случайными компонентами. В нашем примере связь между сезонными и случайными компонентами и колеблемостью практически отсутствует: $r_{u_{случ}u_{сез}} = 0{,}018$, т.е. незначительно отлична от нуля. Показатели общей колеблемости:

$$S(t)_{общ} = \sqrt{\frac{\sum u^2_{общ\,i}}{14-2}} = \sqrt{\frac{83{,}13}{12}} = 2{,}632;$$

$$v(t)_{общ} = \frac{S(t)_{общ}}{\bar{Y}} = \frac{2{,}632}{7{,}78} = 0{,}3383, \text{ или } 33{,}83\%.$$

Общая колеблемость практически равна сезонной.

Общая сумма квадратов отклонений уровней динамического ряда от среднего уровня может быть разложена на три компонента: 1) за счет тренда: $\sum(\hat{Y}_i - \bar{Y}_i)^2$; 2) за счет сезонности: $\sum(\hat{Y}_i' - \hat{Y}_i)$; 3) за счет случайных колебаний: $\sum(Y_i - \hat{Y}_i')^2$. Рассмотрим этот прием на примере динамики реальных доходов населения России[1] (табл. 12.13).

Поскольку тренд индексов реальных доходов населения в отличие от номинальной заработной платы слабый, применим аддитивную модель сезонности. По данным табл. 12.13 получена модель:

$$\hat{Y}' = 66{,}65 + 0{,}5449 \cdot t + 10{,}2 \cdot Z_1 + 14{,}9 \cdot Z_2 + 20{,}4 \cdot Z_3 + 11{,}1 \cdot Z_4 + 19{,}9 \cdot Z_5 + 15{,}7 \cdot Z_6 + 18{,}1 \cdot Z_7 + 17{,}9 \cdot Z_8 + 17{,}4 \cdot Z_9 + 19{,}5 \cdot Z_{10} + 47{,}6 \cdot Z_{11}.$$

Все коэффициенты при Z_j — это средние превышения индекса данного месяца в сравнении с январским уровнем. Переменная Z_j изменяется от Z_1 — для февраля до Z_{11} — для декабря. Все параметры имеют высокую надежность, для модели в целом значение F-критерия составило: $F = 70{,}9$.

[1] Финансовые известия. — 2002. — 24.09.

Таблица 12.13

Анализ динамики реальных доходов населения России (январь 1998 г. = 100%)

Год, мес.	Y_i, %	Тренд $\hat{Y}_i$, %	Модель $\hat{Y}_i'$, %	Отклонения		
				$u_{iобщ}$	$u_{iсез}$	$u_{iобщ}$
1999, 01	72	84	67	−12	−17	5
02	79	85	78	−6	−7	1
03	81	85	83	−4	−2	−2
04	87	86	89	1	3	−2
05	83	87	80	−4	−7	3
06	86	87	90	−1	3	−4
07	84	88	86	−4	−2	−2
08	87	88	89	−1	1	−2
09	87	89	89	−2	0	−2
10	89	89	89	0	0	0
11	90	90	92	0	2	−2
12	124	90	121	34	31	3
2000, 01	73	91	74	−18	−17	−1
02	86	91	84	−5	−7	2
03	93	92	90	1	−2	3
04	95	92	96	3	4	−1
05	90	93	87	−3	−6	3
06	98	93	96	5	7	2
07	94	94	93	0	−1	1
08	95	94	96	1	2	−1
09	96	95	96	1	1	0
10	95	96	96	−1	0	−1
11	101	96	99	5	3	2
12	127	97	127	30	30	0
2001, 01	77	97	80	−20	−17	−3
02	90	98	91	−8	−7	−1
03	96	98	96	−2	−2	0
04	100	99	102	1	3	−2
05	92	99	94	−7	−5	−2
06	105	100	103	5	3	2
07	100	100	99	0	−1	1
08	105	101	102	4	1	3
09	105	101	103	4	2	2
10	104	102	103	2	1	1
11	105	102	105	3	3	0
12	131	104	134	28	31	−3
2002, 01	86	104	87	−18	−17	−1
02	96	104	98	−8	−6	−2
03	102	104	103	−3	−2	−1
04	114	105	109	9	4	5
05	96	106	100	−10	−6	−4
Итого	3896	3896	3896	0	0	0

На основе отклонений, показанных в табл. 12.13, рассчитаны показатели колеблемости:

$$S(t)_{\text{общ}} = \sqrt{\frac{4662}{41-2}} = 10{,}9 \text{ пункта};$$

$$S(t)_{\text{сез}} = \sqrt{\frac{4530}{41-2}} = 10{,}78 \text{ пункта};$$

$$S(t)_{\text{случ}} = \sqrt{\frac{212}{41-2}} = 2{,}33 \text{ пункта}.$$

Сумма квадратов отклонений уровней тренда от среднего уровня равна: $\sum(\hat{Y}_i - Y_i)^2 = 1629$; общая сумма квадратов отклонений исходного ряда от среднего уровня: $\sum(Y_i - \bar{Y}_i) = 6755^2$. Ввиду того, что для изучаемой временно́й выборки различные виды отклонений не являются полностью независимыми, точного разложения сумм по компонентам нет: $6755 \neq 4662 + 1629$, расхождение на 464; сумма квадратов за счет сезонной и случайной колеблемости: $4530 + 212 \neq 4662$, что не равно общей сумме за счет колеблемости, расхождение на 80. Устраним его, умножив слагаемые на отношение 4662:4742, т.е. на 0,983. Получим сумму квадратов отклонений за счет сезонности: $4530 \cdot 0{,}983 = 4583$, за счет случайных колебаний: $212 \cdot 0{,}983 = 209$.

Теперь устраним расхождение с общей суммой квадратов отклонений. Имеем: $4583 + 209 + 1629 = 6421$, тогда как фактически общая сумма квадратов отклонений равна 6755. Следовательно, все компоненты нужно умножить на корректирующий коэффициент, который в нашем примере равен: $6755 : 6421 = 1{,}052$.

Получаем окончательные значения компонентов и их доли в общей сумме квадратов отклонений (табл. 12.14).

Не следует, однако, интерпретировать полученные доли как сравнительные оценки силы влияния разных факторов на развитие явления. Различия уровней за счет колебаний не аккумулируются, и их роль будет «снижаться» тем больше, чем длиннее ряд динамики. К таким оценкам надо подходить очень осторожно.

Таблица 12.14

Разложение общей суммы квадратов отклонений уровня ряда на компоненты

Источники вариаций	Сумма квадратов отклонений	Доля в общей сумме, %
За счет тренда	1714	25,37
За счет сезонности	4821	71,36
За счет случайности	220	3,27
Общая сумма	6755	100

12.11. Прогнозирование на основе тренда и колеблемости

Прогнозирование возможных значений признаков изучаемого объекта — одна из основных задач науки. В ее решении роль статистических методов очень значительна. Одним из них является расчет прогнозов на основе тренда и колеблемости динамического ряда до настоящего времени. Если мы будем знать, как быстро и в каком направлении изменились уровни какого-то признака, то сможем узнать, какого значения достигнет уровень спустя известное время. Методика статистического прогноза по тренду и колеблемости основана на их ***экстраполяции***, т.е. на предположении, что параметры тренда и колебаний сохраняются до прогнозируемого периода. Такая экстраполяция справедлива, если система развивается эволюционно в достаточно стабильных условиях. Чем крупнее система, тем выше вероятность сохранения параметров ее изменения, конечно, на срок не слишком большой! Обычно рекомендуют, чтобы срок прогноза не превышал $^1/_3$ длительности базы расчета тренда.

В отличие от прогноза на основе регрессионного уравнения прогноз по тренду учитывает факторы развития только в неявном виде, и это не позволяет «проигрывать» разные варианты прогнозов при разных возможных значениях факторов, влияющих на изучаемый признак. Зато прогноз по тренду охватывает все факторы, в то время как в регрессионную модель невозможно включить в явном виде более 10—20 факторов в самом лучшем случае.

Сущность прогноза на основе тренда хорошо иллюстрируется следующим рассказом о греческом философе Диогене, жившем в большой бочке на берегу Саронического залива, недалеко от афинского порта Пирея. Как-то вечером Диогена стал окликать снаружи неизвестный. Диоген вышел к нему. «Скажи, мудрый человек, — спросил путник, — дойду ли я к закату в Афины?» Диоген посмотрел на него и сказал: «Иди!» Путник повторил свой вопрос. «Иди!», — закричал Диоген, и путник, пожав плечами, побрел по берегу. «Вернись!», — снова закричал Диоген, и путник вернулся к нему. «Вот теперь я тебе скажу, что до заката ты не дойдешь до Афин. Оставайся у меня». «А почему же ты сразу мне это не сказал, а прогнал меня?» Диоген усмехнулся: «А как же я скажу, дойдешь ли ты до Афин, если я не видел, как быстро ты ходишь?» Прогноз по тренду — это и есть Диогенов прогноз на основе того, как изучаемая система «шла» до настоящего времени.

Рассмотрим методику прогнозирования по тренду с учетом колеблемости на примере цен на нетопливные товары развивающихся стран, тренд и колеблемость которых были измерены в подразд. 12.6 и 12.7 (табл. 12.5 и 12.8). За основу прогнозов возьмем параметры, полученные методом многократного скользящего выравнивания. Параллельно будет показана и методика расчетов при однократном выравнивании.

Итак, имеем уравнение тренда: $\hat{y} = 104{,}53 - 1{,}433t$, где $t = 0$ в 1987 г., оценку генеральной величины среднего квадратического отклонения от тренда $s(t) = 9{,}45$. Эти значения получены при анализе динамики цен весьма значительного сектора мировой торговли, т.е. очень большой и сложной системы. Маловероятно, что условия развития этой системы существенно изменились бы, скажем до 1998 г. Поэтому прогноз на 1998 г. по измеренному тренду можно теоретически считать достаточно обоснованным. Обычно рекомендуется, чтобы период упреждения (от конца базы расчета до прогнозируемого периода) составлял не более $1/3$ длины базы расчета.

Прежде всего вычисляется «*точечный прогноз*» — значение уровня тренда при подстановке в его уравнение номера 1998 г., считая от 1987 г., т.е. $t_k = 11$.

$$\hat{y}_{1988} = 104{,}53 - 1{,}433 \cdot 11 = 88{,}77.$$

Это означает, что наиболее вероятное значение индекса цен на нетопливные товары развивающихся стран в 1998 г. составит около 89% к уровню цен 1990 г., принятому за 100%. Однако параметры тренда, полученные по ограниченному числу уровней ряда, — это лишь выборочные средние оценки, не свободные от влияния распределения колебаний отдельных уровней во времени, как уже сказано ранее. При изменении базы расчета тренда, если, скажем, взять 1977—1993 гг. или 1981—1997 гг., были бы получены несколько иные значения параметров, а значит, и другие значения $\hat{y}_{1998}$. Прогноз должен быть вероятным, как всякое суждение о будущем.

Средняя ошибка прогноза положения линейного тренда на год (момент) с номером t_k вычисляется по формулам.

1. Для однократного выравнивания:

$$m_{\hat{y}_k} = s(t) \cdot \sqrt{\frac{1}{n} + \frac{t_k^2}{\sum t_i^2}}, \qquad (12.39)$$

где t_k — номер года прогноза;

$\sum t_i^2$ — по всей длине ряда n, т.е. $\frac{n^3 - n}{12}$.

2. Для многократного скользящего выравнивания при l сдвигах базы и длине ее n:

$$m_{\hat{y}_k} = s(t) \cdot \sqrt{\frac{1}{n} + \frac{t_k^2}{l \cdot \sum t_i^2}}, \qquad (12.40)$$

где $\sum t_i^2 = \frac{n^3 - n}{12}$.

При $N = 17$, $n = 11$, $l = 7$ получаем:

для однократного выравнивания $m_{\hat{y}_k} = 9{,}133 \cdot \sqrt{\frac{1}{17} + \frac{11^2}{408}} = 5{,}44$;

для многократного выравнивания $m_{\hat{y}_k} = 9{,}45 \cdot \sqrt{\frac{1}{17} + \frac{11^2}{7 \cdot 110}} = 4{,}39$.

Как видим, метод многократного выравнивания на 20% снизил среднюю ошибку прогноза положения тренда.

Для получения достаточно надежных границ прогноза положения тренда, скажем с вероятностью 0,9, следует среднюю ошибку умножить на величину t-критерия Стьюдента при указанной вероятности (или значимости $1 - 0{,}9 = 0{,}1$) при числе степеней свободы, равном для линейного тренда $n - 2$, т.е. 15. Эта величина равна 1,753. Получаем предельную с данной вероятностью ошибку:

$$\alpha = m \cdot t = 439 \cdot 1{,}753 = 7{,}70.$$

Следовательно, с вероятностью 0,9 можно ожидать, что тренд индекса цен в 1998 г. пройдет между значениями $\hat{y}_{1998+\alpha}$ и $\hat{y}_{1998-\alpha}$, т.е. 88,77 + + 7,70 и 88,77 − 7,70; от 81,07 до 96,47 в процентах к уровню цен 1990 г. и, конечно, в одинаковой валюте, без учета ее инфляции.

Однако фактические уровни ряда отклоняются от тренда. Уровень цен в 1998 г. также может быть вовсе не равен уровню положения тренда в этом году. Ошибка прогноза конкретного уровня включает две неопределенности: во-первых, мы не знаем точно, где окажется тренд в 1998 г., а во-вторых, в какую сторону и на сколько уровень ряда отклонится в 1998 г. от положения тренда. Считая, как уже было сказано, колебания случайно (в основном случайно) распределенными во времени, т.е. независимыми от тренда, определим ошибку прогноза уровня конкретного года по правилу сложения независимых дисперсий:

$$m_{y_k} = \sqrt{m_{\hat{y}_k}^2 + s_{(t)}^2}; \qquad (12.41)$$

$$m_{y_k} \sqrt{4{,}39^2 + 9{,}45^2} = 10{,}42.$$

С вероятностью 0,9 ошибка прогноза уровня цен не превзойдет величины 18,27 ($10{,}42 \cdot 1{,}753$), и доверительные границы прогноза составят от 70,5 до 107,0% к уровню 1990 г. Как видим, точность прогноза невелика, разброс возможных значений достиг 37 пунктов, а вероятная ошибка составила 0,206, или 20,6% $\left[\frac{18{,}27}{88{,}77}\right]$ средней величины (точечного прогноза). Можно уменьшить значение вероятной ошибки, если сделать прогноз с меньшей надежностью, скажем с вероятностью 0,75. Тогда значение t-критерия Стьюдента составит 1,197, вероятная ошибка — 12,47 пункта [10,42 · 1,197], доверительные границы — от 76,30 до 101,24% к уровню 1990 г. За уменьшение вероятной ошибки, однако, пришлось заплатить снижением надежности прогноза.

Из имеющейся информации нельзя извлечь больше, чем в ней содержится: как в физике действует закон сохранения массы и энергии, импульса («количества движения»), так здесь действует закон сохранения информации: увеличивая точность, понижаем надежность, увеличивая надежность — понижаем точность. Методика анализа и прогнозирования тоже имеет значение. Она определяет степень полноты извлечения информации, содержащейся в исходном ряду динамики. С помощью методики многократного выравнивания удается более полно извлечь информацию о тренде и уменьшить среднюю ошибку прогноза его положения в прогнозируемом периоде с 5,44 до 4,39. Однако, как видно из (12.41), главной составляющей ошибки прогноза конкретного уровня в нашем расчете является не ошибка прогноза положения тренда, а колеблемость уровней около тренда. Поэтому ошибка прогноза конкретного уровня незначительно сократилась за счет многократного выравнивания. При слабой колеблемости уровней и прогнозировании на значительное удаление от базы главную роль станет играть ошибка положения тренда. Тогда многократное выравнивание даст значительное сокращение средней ошибки прогноза конкретных уровней. Но в любом случае эта ошибка всегда больше показателя колеблемости уровней — среднего квадратического отклонения $S_{y(t)}$[1]. В указанной литературе содержатся формулы для вычисления средней ошибки прогноза положения линии тренда при параболической и экспоненциальной его формах[2]. Если средняя ошибка положения тренда вычислена, ошибку конкретного уровня при любой форме тренда вычисляют по формуле (12.41).

[1] *Четыркин Е. М.* Статистические методы прогнозирования. — М.: Статистика, 1977.

[2] *Юзбашев М. М., Манелля А. И.* Статистический анализ тенденций и колеблемости. — М.: Финансы и статистика, 1983.

12.12. Корреляция рядов динамики

В главах, посвященных статистическому изучению взаимосвязей методом аналитической группировки и методом корреляционного анализа, рассматривались зависимости между признаками, варьирующими в пространственной совокупности. Но необходимо изучать и связи, проявляющиеся в развитии, во времени. Например, есть ли связь между изменениями урожайности сельскохозяйственных культур и изменениями ее себестоимости, рентабельности? Есть ли связь между динамикой рождаемости и динамикой обеспеченности населения жильем? К сожалению, проблема изучения причинных связей во времени очень сложна, и полное решение всех задач такого рода до сих пор не разработано.

Характерным примером для иллюстрации особенностей методики анализа корреляции в рядах динамики служит связь динамики урожайности сельскохозяйственных культур с себестоимостью продукции в 1970—1980-е гг. в СССР. Официально тогда не признавалось наличие инфляции. Однако даже в тех хозяйствах, где применение агротехники прогрессировало и урожайность имела тенденцию роста, себестоимость продукции тоже возрастала. Такой пример представлен в табл. 12.15.

Основная сложность состоит в том, что, как показано в подразд. 12.10, при наличии тренда за достаточно длительный период большая часть суммы квадратов отклонений связана с трендом. Если два признака имеют тренды с одинаковым направлением изменения уровней, то между уровнями этих признаков будет наблюдаться положительная ковариация. И в одном, и в другом ряду уровни более поздних лет будут либо больше, либо меньше уровней более ранних периодов. Коэффициент корреляции уровней окажется положительным. При разной направленности трендов ковариация уровней и коэффициент их корреляции окажутся отрицательными.

Но одинаковая направленность трендов вовсе не означает причинной зависимости. Например, рост производства ракет не причина происходившего в тот же период роста производства мяса. Гораздо вероятнее, что при отсутствии гонки производства ракетного оружия производство мяса росло бы значительно быстрее. А коэффициенты корреляции уровней

Таблица 12.15

Корреляция урожайности картофеля с себестоимостью

Год	Урожайность, ц/га x_i	Себестоимость, руб./ц y_i	Отклонения от средних		$\Delta^2 x_i$	$\Delta^2 y_i$	$\Delta x_i \Delta y_i$	Тренды		Отклонения от трендов		$u_{x_i}^2$	$u_{y_i}^2$	$u_{x_i} u_{y_i}$
			Δx_i	Δy_i				$\hat{x}$	$\hat{y}$	u_{x_i}	u_{y_i}			
1977	108	11,8	−12	−7,2	144	51,84	+86,4	97	+11,7	+11	+0,1	121	0,01	+1,1
1978	81	15,4	−39	−3,6	1521	12,96	+140,4	101	12,9	−20	+2,5	400	6,25	−50,0
1979	106	13,0	−14	−6,0	196	36,00	+84,00	105	14,1	+1	−1,1	1	1,21	−1,1
1980	124	13,9	+4	−5,1	16	26,01	−20,4	108	15,3	+16	−1,4	256	1,96	−22,4
1981	103	15,1	−17	−3,9	289	15,21	+66,3	112	16,6	−9	−1,5	81	2,25	+13,5
1982	106	19,6	−14	+0,6	196	0,36	−8,4	116	17,8	−10	+1,8	100	3,24	−18,0
1983	149	16,2	+29	−2,8	841	7,4	−81,2	120	19,0	+29	−2,8	841	7,84	−81,2
1984	148	17,2	+28	−1,8	784	3,24	50,4	124	20,2	+24	−3,0	576	9,00	−72,0
1985	102	24,0	−18	+5,0	324	25,0	−90,00	128	21,4	−26	+2,6	676	6,76	−67,6
1986	130	22,4	+10	+3,4	100	11,56	+34,0	131	22,7	−1	−0,3	1	0,09	+0,3
1987	80	32,3	−40	+13,3	1600	176,89	−532,0	135	23,9	−55	+8,4	3025	70,56	−462,0
1988	139	24,7	+19	+5,7	361	32,49	+108,3	139	25,1	0	−0,4	0	0,16	0
1989	183	21,4	+63	+2,4	3969	5,76	+151,2	143	26,3	+40	−4,9	1600	24,01	−196,0
Σ	1559	247,0	—	—	10 341	405,16	−111,8	1559	247,0	—	—	7678	133,34	−952,7

высоки! Таким образом, не только возникает масса «ложных корреляций», за которыми нет причинной зависимости, но искажаются (преувеличиваются или преуменьшаются) и те показатели корреляции, за которыми стоят реальные причинные зависимости.

Рассмотрим табл. 12.15. Корреляция уровней урожайности с уровнями себестоимости картофеля отсутствует: коэффициент корреляции равен −0,055, т.е. незначимо отличен от нуля. Но ведь на самом деле по законам экономики при пространственной корреляции в совокупности хозяйств связь урожайности и себестоимости сильная, обратная — чем выше урожайность, тем ниже себестоимость.

Среднее значение урожайности по данным табл. 12.15 составило: $\bar{x} = 119{,}92$ ц/га, средняя себестоимость: $\bar{y} = 19{,}0$ руб./ц. Уравнения трендов:

урожайности $\hat{x} = 119{,}9 + 3{,}81t$;

себестоимости $\hat{y} = 19{,}0 + 1{,}22t$

при $t = 0$ в 1983 г.

Всесторонний экономический и статистико-математический анализ ситуации показывает, что причина отсутствия корреляции уровней в том, что оба признака имеют одинаково направленные тренды — возрастание урожайности происходило параллельно с возрастанием себестоимости, вовсе не являясь причиной последнего! Себестоимость росла из-за инфляции в стране, влияние которой оказалось сильнее, чем направленное на снижение себестоимости влияние роста урожайности.

Если же рассматривать уровни признаков год за годом, легко заметить, что снижению урожайности в сравнении с предыдущим годом соответствовал рост себестоимости, а повышению урожайности — ее снижение, т.е. связь обратная, которая и должна быть. Следовательно, чтобы получить реальные показатели корреляции, необходимо абстрагироваться от искажающего влияния трендов: вычислить отклонения уровней урожайности и себестоимости от трендов и измерить корреляцию не уровней, а колебаний двух признаков.

Подставляя в формулу парного коэффициента корреляции (9.11) вместо уровней признаков их отклонения от трендов u_{x_i}, u_{y_i}, получаем:

$$r_{u_x u_y} = \frac{\sum\limits_{(i)} (u_{x_i} - \bar{u}_x)(u_{y_i} - \bar{u}_y)}{\sqrt{\sum\limits_{(i)} (u_{x_i} - \bar{u}_x)^2 \sum\limits_{(i)} (u_{y_i} - \bar{u}_y)^2}}. \quad (12.42)$$

Однако среднее отклонение от тренда равно нулю (для прямой и параболы всегда, а для других типов тренда лишь в том случае, если правильно отражают тенденцию), $\bar{u}_x = \bar{u}_y = 0$. Подставив в (12.42) $\bar{u}_y = 0$ и $\bar{u}_x = 0$, получим:

$$r_{u_x u_y} = \frac{\sum\limits_{(i)} u_{x_i} u_{y_i}}{\sqrt{\sum\limits_{(i)} u_{x_i}^2 \sum\limits_{(i)} u_{y_i}^2}}. \quad (12.43)$$

Коэффициент регрессии для линейной зависимости принимает вид:

$$b = \frac{\sum\limits_{(i)} u_{x_i} u_{y_i}}{\sum\limits_{(i)} u_{x_i}^2}. \quad (12.44)$$

Свободный член линейного уравнения регрессии

$$a = \bar{u}_y - b\bar{u}_x = 0.$$

Уравнение регрессии отклонений от тренда имеет вид:

$$\tilde{u}_y = bu_x. \quad (12.45)$$

По данным табл. 12.15 коэффициент корреляции уровней урожайности и себестоимости

$$r_{x,\,y} = \frac{-111{,}8}{\sqrt{10341 \cdot 405{,}16}} = -0{,}055.$$

Прямая связь одинаково направленных трендов почти полностью компенсировала обратную связь между колебаниями признаков. Из тринадцати произведений $\Delta_{x_i} \Delta_{y_i}$ семь положительны. Прежде всего в начале и в конце ряда, где сильнее всего сказались тренды. Если бы не страшный неуро-

жай в 1987 г., вызвавший огромные отклонения уровней, коэффициент корреляции был бы даже положителен.

Напротив, корреляция отклонений от трендов дает результат, соответствующий экономическому содержанию связи урожайности с себестоимостью.

Коэффициент корреляции отклонений от трендов по формуле (12.43) составил:

$$r_{u_x u_y} = \frac{-952{,}7}{\sqrt{7678 \cdot 133{,}3}} = -0{,}941.$$

Коэффициент детерминации $r^2_{u_x u_y}$ равен 0,88, или 88% колебаний себестоимости картофеля связаны с колебаниями урожайности. Положительны лишь три произведения отклонения $u_{x_i} \cdot u_{y_i}$, притом наименьшие.

Коэффициент регрессии по формуле (12.44)

$$b = \frac{-952{,}7}{7678} = -0{,}124.$$

Уравнение регрессии:

$$\tilde{u}_{y_i} = -0{,}124\, u_{x_i}.$$

Это означает, что в среднем за период отклонение себестоимости от тренда было противоположно по знаку и составляло 0,124 отклонения урожайности от своего тренда. Если, например, урожайность в 1993 г. окажется на 20 ц/га ниже уровня тренда для этого года, составляющего: $119{,}9 + 3{,}81 \cdot 10 = 158$ ц/га, то себестоимость надо ожидать на $20(-0{,}124) = 2{,}48$ руб. за 1 ц выше уровня тренда, который для 1993 г. равен 31,2 руб за 1 ц, т.е., учитывая и тренды, и предполагаемый плохой урожай в 1993 г., себестоимость картофеля составила бы: $31{,}2 + 2{,}48 = 33{,}66$ руб./ц. Естественно, что этот прогноз всего лишь пример, как пользоваться уравнением регрессии отклонений от тренда. В нашем случае метеорология не дает оснований для прогноза урожайности, а сильная инфляция делает вообще невозможным любой прогноз себестоимости без использования дефлятора.

Данные табл. 12.15 позволяют сделать интересное заключение о различии характера динамики признаков. Если из общей дисперсии (суммы квадратов отклонений от среднего

уровня) урожайности 10 341 большую часть составляет дисперсия за счет колеблемости 7678, то для себестоимости преобладающим моментом общей дисперсии, равной 405,16, является не колеблемость, дающая только 133,43, а тренд; это эффект скрытой инфляции до 1989 г.

Другим приемом измерения корреляции в рядах динамики может служить корреляция между теми из цепных показателей рядов, которые являются константами из трендов. При линейных трендах — это цепные абсолютные приросты. Вычислив их по исходным рядам динамики (a_{xi}, a_{yi}), находим коэффициент корреляции между абсолютными изменениями по формуле (12.43) или, что более точно, по формуле (12.42), так как средние изменения не равны нулю в отличие от средних отклонений от трендов. Допустимость данного способа основана на том, что разность между соседними уровнями в основном состоит из колебаний, а доля тренда в них невелика, следовательно, искажение корреляции от тренда очень большое при кумулятивном эффекте на протяжении длительного периода, весьма мало — за каждый год в отдельности. Однако нужно помнить, что это справедливо лишь для рядов с c-показателем, существенно меньшим единицы. В нашем примере для ряда урожайности c-показатель равен 0,144, для себестоимости он равен 0,350. Коэффициент корреляции цепных абсолютных изменений составил 0,928, что очень близко к коэффициенту корреляции отклонений от трендов.

Для рядов с тенденцией, близкой к экспоненте, следует рекомендовать корреляцию цепных темпов роста. Вычисление корреляции рядов динамики по цепным показателям не требует предварительного вычисления трендов, но все же желательно иметь приближенное представление о характере тенденции. Для параболических трендов с не очень большими ускорениями можно коррелировать цепные абсолютные изменения; при больших ускорениях лучше их не коррелировать. Если коррелируемые ряды имеют разные типы тенденций, вполне допустимо коррелировать соответствующие разные цепные показатели: абсолютные изменения в одном ряду с темпами изменений в другом и т.д.

К сожалению, все вышеизложенные приемы, по существу, решают только задачу измерения связи между колебаниями признаков, а не между тенденциями их изменений. Насколь-

ко допустимо переносить выводы о тесноте связи между колебаниями на связь динамических рядов в целом, зависит от материального, качественного содержания процесса и причинного механизма связи. Это проблема, выходящая далеко за пределы статистической науки. Если колебания урожайности являются на самом деле следствиями колебания суммы осадков за лето, т.е. корреляция именно колебаний отвечает сущности причинной связи, то, например, причинную связь между дозой удобрений и урожайностью нельзя свести к зависимости только между колебаниями. Здесь главное — причинная связь тенденций, а измерять ее мы так и не научились.

Завершая этим признанием главу о статистическом анализе рядов динамики, дадим последние методологические советы изучающим статистику.

Всякая наука — это процесс продолжающегося познания природы и общества. Нет наук законченных, которые следует лишь выучить наизусть, чтобы все знать. Учебники и учебные пособия — лишь сжатые и неполные изложения уже достигнутого наукой уровня познания. Изучайте специальную литературу, если хотите больше знать, а также новейшие достижения ученых всего мира.

Не считайте и себя только «сосудами для вливания» знаний. Познав известное, вы можете (и должны!) внести свой вклад в дальнейшее развитие теории статистики. «Если не я, то кто же?»

РЕЗЮМЕ

Динамический ряд включает значения показателя за последовательные периоды или моменты времени. Каждое значение показателя называется уровнем ряда.

Динамика показателя может включать тенденцию и колебания (отклонения от тенденции). Колебания могут быть регулярными (циклическими), в том числе сезонными, и нерегулярными (случайными). Тенденция динамики связана с действием долговременных причин и условий развития. Колебания связаны с действиями краткосрочных или циклических факторов.

Тенденция и колебания хорошо видны на графике.

Изменения уровней временного ряда характеризуют абсолютные и относительные показатели динамики: абсолютный прирост (цепной и базисный), ускорение абсолютного изменения, темп роста (цепной и базисный), темп роста, абсолютное значение одного процента прироста.

Средний уровень динамического ряда рассчитывается по формуле средней арифметической простой (для интервального ряда) либо взвешенной (для моментного ряда) и используется для обобщенной характеристики периода развития, для сравнения средних достижений в разные периоды. Средний темп динамики рассчитывается по формуле средней геометрической. Средний абсолютный прирост определяется по формуле средней арифметической. Расчет среднегодового темпа динамики, требуемого для достижения заданного уровня, проводится по формуле А. и И. Соляников.

При параболической тенденции среднегодовые темпы легко получить, пользуясь таблицей, составленной Л. С. Казинцом [4].

При анализе динамики важно оценивать продолжительность срока, за который один объект («отстающий») может догнать другой объект («передовой»).

Для выявления тренда нужно решить: были ли условия развития достаточно однородными, каков характер действия основных факторов развития?

Среди основных форм тренда выделяются: линейный, параболический, экспоненциальный, логарифмический, тренд в форме степенной функции, гиперболы, логистической форме.

Для выявления тенденции и устранения колебаний можно воспользоваться методом скользящей средней.

Параметры уравнения тренда находятся МНК. При этом может быть использован метод условного нуля, т.е. центральный член ряда принимается за точку отсчета. Уравнение тренда $\hat{y} = a + bt$, полученное при этом, будет отличаться от уравнения тренда, полученного при значениях $t = 1, 2, ..., n$, только свободным членом a, а значения параметра b будут одинаковы в обоих уравнениях.

При вычислении параметров тренда уровни исходного ряда входят с разными весами — значениями t. Поэтому влияние колебаний уровней на параметры тренда зависит от того, на какой год приходится либо высокое, либо низкое значе-

ние. Для полного исключения влияния колебаний на параметры тренда следует применять метод многократного скользящего выравнивания. Установить тип колеблемости (пилообразная, или маятниковая, долгопериодическая циклическая, случайно распределенная по времени) можно с помощью критерия поворотных точек Кендэла.

Интенсивность колеблемости измеряют с помощью следующих показателей: среднего линейного отклонения от тренда, среднего квадратического отклонения от тренда, коэффициента колеблемости.

Анализ типа колеблемости и определение длины цикла могут быть основаны на расчете коэффициентов автокорреляции отклонений от тренда.

Оценку степени устойчивости реализации тренда можно провести с помощью коэффициента корреляции рангов Спирмена. Устойчивость тренда может быть измерена соотношением между среднегодовым абсолютным изменением и среднеквадратическим отклонением уровней от тренда.

Анализ сезонности проводится на основе анализа дисперсии уровней временно́го ряда. Выделяется дисперсия за счет тренда, за счет сезонных колебаний, за счет случайных колебаний (остаточная). Графически сезонность изображается либо в виде сезонной волны, либо в виде радиальной диаграммы.

При высокой надежности уравнения тренда оно может использоваться для прогнозирования уровней временно́го ряда (с учетом сезонной компоненты). Следует иметь в виду, что средняя ошибка прогноза всегда превышает показатель колеблемости уровней.

При изучении взаимосвязи между динамикой разных показателей следует опасаться неверных умозаключений, вызываемых ложной корреляцией, поскольку все показатели изменяются с изменением времени t, которое может рассматриваться в качестве общей причины для всех временных рядов. Для того чтобы устранить ложную корреляцию, рассчитывают коэффициент корреляции не между уровнями временных рядов, а между отклонениями от тренда или первыми разностями при наличии линейных трендов.

Уравнение регрессии, описывающее зависимость динамики одного показателя от другого, строится либо по отклонениям от тренда, либо по первым разностям (в случае линейных трендов), либо по уровням временны́х рядов при включении переменной «время» t в уравнение в качестве объясняющей переменной.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Андерсон Т.* Статистический анализ временных рядов: Пер. с англ. — М.: Мир. —1976.
2. *Афанасьев В. Н., Юзбашев М. М.* Анализ временных рядов и прогнозирование. — М.: Финансы и статистика, 2001.
3. *Вайну Я. Я.-Ф.* Корреляция рядов динамики. — М.: Статистика, 1977.
4. *Казинец Л. С.* Темпы роста и абсолютные приросты. — М.: Статистика, 1975.
5. *Четыркин Е. М.* Статистические методы прогнозирования. — 2-е изд. — М.: Финансы и статистика, 1983.
6. *Юзбашев М. М., Манелля А. И.* Статистический анализ тенденций и колеблемости. — М.: Финансы и статистика, 1983.

Глава 13

ИНДЕКСЫ

13.1. Понятие индекса

Само слово «индекс» (index) означает показатель. Обычно этот термин употребляется для некой обобщающей характеристики изменений. Например, уже знакомый вам индекс Доу-Джонса, индекс деловой активности, индекс объема промышленного производства и т.д. Гораздо реже термин «индекс» употребляется как обобщенный показатель состояния, например, известный коэффициент умственного развития *IQ*.

В этой главе мы рассмотрим индексы прежде всего как показатели изменений. Очевидно, что сфера использования таких показателей безгранична: спортсмены стремятся улучшить свои достижения, предприниматель желает увеличить прибыль и т.д. Во всех этих случаях необходимо выразить изменения количественно. Как изменились цены, уровень жизни, покупательная сила денег и прочее? Ответы на все эти вопросы позволяют дать индексы.

В предыдущей главе вы познакомились с показателями, которые измеряют абсолютные и относительные изменения: темпы роста, прироста, абсолютный прирост, цепные и базисные показатели, показатели средних изменений за период. В чем же специфика индексов? Принципиальных отличий три.

Во-первых, индексы позволяют измерить изменение сложных явлений. Например, нужно определить: как изменились за год расходы жителей Москвы на городской транспорт? Для ответа на этот вопрос вы должны иметь численность пассажиров, перевезенных за год каждым видом городского транспорта, рассчитать среднемесячную численность

пассажиров или взять точные данные из отчетов по месяцам, умножить численность на тариф перевозки (и число месяцев его действия в случае использования среднемесячной численности) и полученные величины просуммировать. То же нужно сделать по данным за прошлый год. Затем сопоставить сумму расходов за последний год с суммой за прошлый год, т.е. это не просто сравнение двух чисел, как при расчете темпов динамики или приростов, а получение и сравнение некоторых агрегированных величин.

Во-вторых, индексы позволяют проанализировать изменение — выявить роль отдельных факторов. Например, можно определить, как изменилась сумма выручки городского транспорта за счет изменения численности пассажиров и тарифов, наконец, за счет соотношения в объеме перевозок разными видами транспорта.

В-третьих, индексы являются показателями сравнений не только с прошлым периодом (сравнение во времени), но и с другой территорией (сравнение в пространстве), а также с нормативами. Например, интересно знать, не только как изменилось среднедушевое потребление мяса в России в данном году по сравнению с прошлым годом (или с каким-либо другим периодом), но и сравнить показатели среднедушевого потребления мяса в России и развитых странах Запада, Востока, а также провести сравнение с нормативной величиной, отвечающей нормам рационального питания. Очевидно, что каждое направление сравнения вносит что-то новое. Так, доля расходов на фундаментальные исследования и содействие научно-техническому прогрессу в России в 2002 г. составила в процентах к ВВП 1,56%. Это меньше, чем было в 2001 г., когда эта доля составляла 1,85%. Сравнение показателей 2002 г. и 2001 г. показывает снижение на 16 процентных пунктов (1,56 : 1,85 = 0,84). Если же сравнить данные России с данными стран ОЭСР, где инвестиции в фундаментальные исследования и содействие научно-техническому прогрессу в 2002 г. составляли 4,7% от ВВП, то результаты будут еще менее оптимистичными — соответствующий индекс составляет: 1,56 : 4,7 = 0,33, или 33 процентных пункта.

Существует несколько определений индекса. Приведем одно из них, может быть самое краткое.

Индекс — это показатель сравнения двух состояний одного и того же явления (простого или сложного, состоящего из соизмеримых или несоизмеримых элементов).

Каждый индекс включает два вида данных: оцениваемые данные, которые принято называть *отчетными* и обозначать значком «1», и данные, которые используются в качестве базы сравнения, — *базисные*, обозначаемые значком «0».

Индекс, который строится как сравнение обобщенных величин, называется *сводным*, или *общим*, и обозначается *I*. Если же сравниваются необобщенные величины, то индекс называется *индивидуальным* и обозначается *i*. Как правило, подстрочно дается значок, который указывает, для оценки какой величины построен индекс. Например, $I_{q1/0}$ или $i_{q1/0}$, т.е. сводный и индивидуальный индекс для величины q.

Сравнения во времени могут охватывать короткий период: выработка за текущий и за вчерашний день, цены в сентябре по сравнению с августом и т.д. Но сравнение может проводиться и с отдаленным периодом: современные данные с довоенным 1940 г. или с 1986 г. — годом начала перестройки, когда экономика еще не была затронута структурными изменениями и т.д. Выбор базисного периода всегда аргументирован той задачей, для которой строится индекс. Обычно руководствуются двумя правилами: либо база сравнения представляет стабильное состояние, либо экстремальное значение — высшее достижение или низший уровень (в случае падения экономических показателей). Конечно, сравнение с отдаленным периодом вносит дополнительные трудности, что уже отмечалось в гл. 12. Некоторые специфические для построения индексов проблемы будут затронуты ниже.

13.2. Индекс как показатель центральной тенденции (индекс средний из индивидуальных)

Вы можете услышать, что уровень потребительских цен понизился или повысился. Речь в этом случае идет об индексе цен на потребительские товары. Общее изменение образуется под влиянием изменений цен на отдельные товары. Таким образом, мы имеем ряд отношений:

$$\frac{p_{11}}{p_{01}}, \frac{p_{12}}{p_{02}}, \frac{p_{13}}{p_{03}}, \ldots, \frac{p_{1n}}{p_{0n}} \text{ и т.д.}$$

Эти отношения не что иное, как индивидуальные индексы, и сводный индекс представляет собой средний из них:

$$I_p = \overline{\left(\frac{p_{1j}}{p_{0j}}\right)},$$

где j — номер товара.

Поскольку средняя — показатель центра распределения, то и сводный индекс можно назвать показателем центральной тенденции. Проблема заключается в том, как получить этот сводный индекс. Впервые она возникла при попытке оценить совокупное изменение цен либо в виде отношения сумм цен:

$$\frac{p_{11} + p_{12} + \ldots + p_{1n}}{p_{01} + p_{02} + \ldots + p_{0n}} = \frac{\sum_{j=1}^{n} p_{1j}}{\sum_{j=1}^{n} p_{0j}},$$

либо как среднее из изменений цен на отдельные товары:

$$\frac{\frac{p_{11}}{p_{01}} + \frac{p_{12}}{p_{02}} + \ldots + \frac{p_{1n}}{p_{0n}}}{n} = \frac{\sum_{j=1}^{n} \frac{p_{1j}}{p_{0j}}}{n}. \quad (13.1)$$

В том и другом варианте представлены невзвешенные средние. Первое решение основано на том, что цена рассчитывается за единицу товара, например за 1 кг, и сумма цен может рассматриваться как набор слагаемых с равными весами. Однако этот вариант не отвечает задаче осреднения показателей изменений цен на отдельные товары. Второй вариант настораживает тем, что согласно общему правилу средняя из относительных величин должна вычисляться как средняя взвешенная. Действительно, если говорить конкретно об измерении динамики цен на все продовольственные или непродовольственные товары, то ясно, что если цены на ювелирные изделия из золота удвоятся, а цены на хлеб останутся неизменными, это не значит, что в целом цены выросли на 50% ((2 + 1)/2 = 1,5). Приведенный пример показывает, что ин-

декс цен для каждого товара должен сопровождаться неким «весом», который позволяет оценить относительную значимость этого индекса для потребителя. В качестве веса используют удельный вес в общей стоимости покупок в базисном периоде:

$$d_0 = \frac{q_{0j}p_{0j}}{\sum\limits_{(i)} q_{0j}p_{0j}}.$$

Если обозначить удельный вес отдельных затрат d_{0i}, то получим общий индекс цен как средний арифметический взвешенный из индивидуальных индексов цен:

$$I_p = \frac{\sum\limits_{(j)} i_{pj}d_{0j}}{\sum\limits_{(j)} d_{0j}} = \frac{\sum\limits_{(j)} i_{pj}q_{0j}p_{0j}}{\sum\limits_{(j)} q_{0j}p_{0j}}, \qquad (13.2)$$

т.е. $I_p = \bar{i}_p$.

Используя формулу (13.2), можно получить общее изменение цен на продукты по данным табл 13.1.

Распространено утверждение, что чем сильнее варьируют веса средней, тем значительнее отличие невзвешенной средней от взвешенной. Покажем ошибочность этого утверждения применительно к индексу среднему из индивидуальных.

Пример А. Равенство взвешенной и простой средних при сильной вариации весов.

В табл. 13.1 представлены данные примера А.

Таблица 13.1

Данные примера А

Номер товара	Цена		Индекс i_p	Доля в базисной выручке d_0	$i_p d_0$	Вариация долей	
	P_0	P_i				$dj_0 - \bar{d}_0$	$(d_{j0} - \bar{d}_0)^2$
1	10	11	1,1	0,40	0,44	0,20	0,0400
2	15	30	2,0	0,25	0,50	0,05	0,0025
3	20	28	1,4	0,15	0,21	−0,05	0,0025
4	25	40	1,6	0,10	0,16	−0,10	0,0100
5	30	27	0,9	0,10	0,09	−0,10	0,0100
Итого			1,4	1,00	1,40	0	0,0650

Невзвешенный средний индекс цен $\bar{i}_p = \frac{\sum i_p}{5} = \frac{7}{5} = 1{,}4.$

Среднее значение веса $\bar{d}_0 = \frac{1}{5} = 0{,}2.$

Взвешенный средний индекс цен $\bar{i}_p = \frac{\sum i_p d_0}{\sum d_0} = \sum i_p d_0 = 1{,}4.$

Результат совпадает с простой средней. Между тем вариация весов значительна, стандартное отклонение

$$s_d = \sqrt{\frac{0{,}0650}{5}} = 0{,}1140175.$$

Коэффициент вариации весов

$$v_d = \frac{0{,}1140175}{0{,}2} = 0{,}5700, \text{ т.е. } 57\%.$$

Пример Б. Неравенство взвешенной и простой средних при слабой вариации весов.

В табл. 13.2 представлены данные примера Б.

Невзвешенный средний индекс цен $\bar{i}_p = \frac{7}{5} = 1{,}4.$

Взвешенный средний индекс цен $\bar{i}_p = \frac{1{,}486}{1} = 1{,}486.$

Вариация весов $s_d = \sqrt{\frac{0{,}0112}{5}} = 0{,}04733,$

Таблица 13.2

Данные примера Б

Номер товара	Цена		Индекс i_p	Доля в базисной выручке d_0	$i_p d_0$	Вариация долей	
	P_0	P_i				$dj_0 - \bar{d}_0$	$(d_{j0} - \bar{d}_0)^2$
1	10	11	1,1	0,15	0,165	−0,05	0,0025
2	15	30	2,0	0,26	0,520	0,06	0,0036
3	20	28	1,4	0,19	0,266	−0,01	0,0001
4	25	40	1,6	0,25	0,400	0,05	0,0025
5	30	27	0,9	0,15	0,135	−0,05	0,0025
Итого	x	x	1,4	1,00	1,486	0	0,0112

$v_d = 0{,}2366$, или 23,7%, т.е. вариация весов намного слабее, чем в примере А.

Рассмотрим: в чем секрет таких соотношений. Обратимся к формуле взвешенной средней:

$$\bar{x}_{\text{взвеш}} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} (\bar{x} + \Delta x_i) \cdot (\bar{f} + \Delta f_i)}{\sum_{i=1}^{n} (\bar{f}_i + \Delta f_i)},$$

где $x, \bar{f}$ — простые средние;
$\Delta x, \Delta f$ — отклонения от них.

Представим последнее выражение как:

$$\frac{\sum \bar{x}\bar{f} + \sum \bar{x}\Delta f_i + \sum \bar{f}\Delta x_i + \sum \Delta x_i \Delta f_i}{\sum \bar{f} + \sum \Delta f_i} =$$

$$= \frac{n\bar{x}\bar{f} + \bar{x}\sum \Delta f_i + \bar{f}\sum \Delta x_i + \sum \Delta x_i \Delta f_i}{n\bar{f} + 0} =$$

$$= \frac{n\bar{x}\bar{f} + \bar{x} \cdot 0 + \bar{f} \cdot 0 + \sum \Delta x_i \Delta f_i}{n\bar{f}} = \bar{x} + \frac{\sum \Delta x_i \Delta f_i}{n\bar{f}}.$$

Числитель второго слагаемого можно представить через коэффициент корреляции между x и f:

$$\bar{x}_{\text{взвеш}} = \bar{x} + \frac{n\sigma_x \sigma_f r_{xf}}{n\bar{f}} = \bar{x} + \sigma_x \cdot v_f \cdot r_{xf}. \qquad (13.3)$$

Эта формула аналогична формуле (5.6).

Следовательно, средняя взвешенная равна простой средней, если:

- вариация признака x_i отсутствует, т.е. $\sigma_x = 0$;
- вариация весов f_i отсутствует, т.е. $v_f = 0$;
- нет корреляции между вариациями признака и весов, т.е. $r_{xf} = 0$ (хотя сами x_i и f_i варьировали как угодно сильно).

Отношение взвешенной средней и простой можно выразить следующим образом:

$$\frac{\bar{x}_{\text{взвеш}}}{\bar{x}_{\text{прост}}} = 1 + v_x \cdot v_f \cdot r_{xf}. \qquad (13.4)$$

Поскольку различие взвешенной и простой средних зависит от корреляции значений признака и веса, оно может оказаться большим при слабой вариации весов, чем при из сильной вариации (см. гл. 5).

Рассмотрим соотношение между индексами (13.1) и (13.2) на примере табл. 13.3.

Таблица 13.3

Данные розничной торговли города N

	Выручка в мае		Отношение цен в июне к ценам в мае, % $i_p = p_1 : p_0$	Выручка с учетом изменения цен, млн руб. $q_0 p_1 = q_0 p_0 \cdot i_p$
	абс., млн руб.	относит.		
	$q_0 p_0$	d_0		
1	2	3	4	5
Мясо и мясопродукты	2352,0	0,271	110,5	2599,0
Рыба и рыбопродукты	735,0	0,085	112,2	824,7
Масло животное	2058,0	0,237	103,2	2123,8
Масло растительное	9,8	0,001	105,6	10,4
Молоко и молочные продукты	882,0	0,102	102,4	903,2
Сахар	2644,0	0,304	107,3	2837,0
Итого	8680,8	1,000	641,2*	9298,1

* Обычно i_p не суммируются.

Обратите внимание на данные графы 5 табл. 13.3: произведение $q_0 p_0 \cdot i_p$ имеет не просто техническое значение взвешивания индивидуального индекса, но и дает определенный содержательный результат — показатель условных затрат на покупку базисных товаров (q_0) с учетом изменения цен: $q_0 \cdot p_0 \cdot i_p = = q_0 \cdot p_1$.

Это дает право представить формулу (13.2) в виде

$$I_p = \frac{\sum_{(j)} q_{0j} p_{1j}}{\sum_{(j)} q_{0j} p_{0j}}. \qquad (13.5)$$

Выражение (13.5) получило известность как индекс Ласпейреса, предложившего эту формулу в 1864 г.

По данным табл. 13.3

$$I_p = \frac{9298{,}1 \text{ млн руб.}}{8680{,}8 \text{ млн руб.}} = 100\% = 107{,}1\%,$$

т.е. цены возросли в среднем на 7,1%. Если воспользоваться формулой (13.1), то $I_p = 641{,}2/6 = 1{,}069 \cdot 100 = 106{,}9\%$, т.е. в среднем цены возросли на 6,9%. Отличие от среднего взвешенного арифметического индекса составляет 0,2%.

Мы рассмотрели определение среднего изменения на основе средней арифметической из индивидуальных, но ведь могут использоваться и другие виды средних: средняя геометрическая, средняя гармоническая и т.д. — невзвешенные и взвешенные. Используя среднюю геометрическую невзвешенную, получаем:

$$I_p = \sqrt[6]{1{,}105 \cdot 1{,}122 \cdot 1{,}032 \cdot 1{,}056 \cdot 1{,}024 \cdot 1{,}073} = \sqrt[6]{1{,}486} \cdot 100\% = 106{,}8\%.$$

Средняя гармоническая всегда дает результат, меньший средней арифметической. Применяя среднюю гармоническую невзвешенную, получаем:

$$I_p = \frac{6}{\frac{1}{1{,}105} + \frac{1}{1{,}122} + \frac{1}{1{,}032} + \frac{1}{1{,}056} + \frac{1}{1{,}024} + \frac{1}{1{,}073}} = 1{,}068 \cdot 100\% = 106{,}8\%.$$

Опять-таки деление единицы на каждый индекс предполагает равное значение изменения цен на товары, что не соответствует практике.

Используя в качестве весов затраты на покупку в отчетном периоде, получаем сводный индекс цен как средний гармонический взвешенный из индивидуальных индексов цен:

$$I_p = \frac{\sum q_1 p_1}{\sum \frac{q_1 p_1}{i_p}}. \qquad (13.6)$$

В формуле (13.6) и далее для простоты мы опустили подстрочный значок, соответствующий номеру товара (элемента), хотя, конечно же, суммирование и в числителе, и в знаменателе проводится по всему набору товаров (элементов).

Таблица 13.4

Данные розничной торговли города

№ п/п	Относительное изменение количества купленных продуктов в июне по сравнению с маем, %, $i_p = q_1 : q_0$	Выручка в июне, млн руб. $q_1 p_1$	Условная выручка без учета изменения цен, млн руб. $q_1 p_0 = q_1 p_1 : i_p$
1	98,5	2560,0	2316,7
2	100,3	827,2	737,3
3	97,8	2077,1	2012,7
4	102,0	10,6	10,0
5	100,0	903,2	882,0
6	98,0	2780,3	2591,1
Итого	596,6*	9158,4	8549,8

* Обычно i_q не суммируются.

Рассчитаем этот индекс по данным табл. 13.3. Кроме того, нам потребуются дополнительные данные. Как всегда, лучшей формой представления цифровых данных является таблица. Поместим все необходимые данные в табл. 13.4, используя вместо названий номера продуктов.

$$I_p = \frac{9158{,}4 \text{ млн руб.}}{8549{,}8 \text{ млн руб.}} \cdot 100\% = 107{,}1\%.$$

Результат совпал с тем значением I_{p1}, которое было получено по формуле (13.2). Но это случайное совпадение, которое оказалось возможным из-за слабой корреляции между изменением уровня цен и объемом продаж отдельных товаров. Подобное может быть при сравнении за короткий период. В рыночной экономике взаимосвязь между колебаниями цен и объемом продаж проявляется при сравнении за более длительный период. Ниже будет показано, как измерить величину этой корреляции (см. формулу (13.7)).

Знаменатель формулы (13.6) имеет смысл затрат на покупку «отчетного» количества товаров по базисным ценам:

$$\left(\frac{q_1 \cdot p_1}{\frac{p_1}{p_0}}\right) = q_1 p_0$$

Тогда формула (13.6) может быть представлена как

$$I_p = \frac{\sum\limits_{(J)} q_1 p_1}{\sum\limits_{(J)} q_1 p_1}. \qquad (13.7)$$

Эта формула индекса цен была предложена Г. Пааше в 1874 г. Различие между индексами Пааше и Ласпейреса, их использование обсуждаются ниже.

Итак, мы рассмотрели применение разных форм и видов средних величин для определения среднего изменения цен по всем товарам. Люди всегда в первую очередь интересовались ценами и их изменениями. Но такой же подход может быть применен к оценке сводных изменений других характеристик, например объема (количества) покупок товаров. Кстати, заметим, что используемые нами обозначения цен (p), количества (q) неслучайны и соответствуют начальным буквам английских слов price (цена) и quantity (количество). Это закрепленные обозначения в статистике.

Таким образом, общее изменение количества проданных товаров формируется как среднее по отношению к изменениям объема покупок отдельных товаров, т.е.

$$I_q = \bar{i}_q,$$

где $i_q = \frac{q_1}{q_0}$.

Возникает вопрос о порядке расчета средней из i_q: средняя арифметическая — простая или взвешенная — или другая форма средней. Ограничимся рассмотрением только средней арифметической.

По данным табл. 13.4 простая средняя арифметическая из индивидуальных индексов количества равна:

$$I_q = \frac{\sum\limits_{(J)} i_q}{n} = \frac{596{,}6}{6} = 0{,}994 \cdot 100\% = 99{,}4\%(-0{,}6\%).$$

Используя в качестве весов для изменений объема покупок удельный вес покупок в общей сумме затрат, получаем:

$$I_p = \frac{\sum\limits_{(J)} i_q \cdot d_0}{\sum\limits_{(J)} d_0} = \frac{\sum\limits_{(J)} i_q \cdot q_0 \cdot p_0}{\sum\limits_{(J)} q_0 \cdot p_0}, \qquad (13.8)$$

т.е. индекс I_q — средний арифметический взвешенный из индивидуальных i_q.

По данным нашего примера (табл. 13.3 и 13.4) общий индекс количества равен:

$$I_q = \frac{8549{,}8 \text{ млн руб.}}{8680{,}8 \text{ млн руб.}} \cdot 100\% = 98{,}5\%.$$

Получилось, что объем покупок продовольственных товаров сократился в среднем на 1,5%. Это более значительная оценка снижения, нежели полученная при расчетах с простой средней арифметической (−0,6%). Так что мы еще раз получили подтверждение зависимости результата от использованной формулы.

Зная среднюю величину изменения показателя и индивидуальные индексы, можно проводить анализ методами вариационной статистики: анализировать распределение товаров по изменению цен, объема покупок, сравнивать модальное и среднее изменение, максимальное и минимальное; по показателям эксцесса распределений делать выводы о том, насколько однородны изменения цен и количества по отдельным товарам, группировать товары по уровню цен и степени их изменения и т.д.

13.3. Агрегатные индексы. Система индексов

Мы познакомились с построением сводных индексов на основе индивидуальных. Однако возможен и другой путь. Обратимся к формулам Ласпейреса (13.5) и Пааше (13.7). Эти индексы могут быть рассчитаны на основе данных о количестве проданных товаров в базисном и отчетном периодах (по каждому j-му товару) q_{0j} и q_{1j} и ценах — p_{1j} и p_{0j}. Такие индексы принято называть агрегатными. Так же можно постро-

ить и I_q: не через осреднение индивидуальных индексов, а на основе сравнения двух сумм (агрегатов).

Агрегатные индексы считаются основной формой индексов. Они выполняют две функции: *синтетическую* и *аналитическую*. Первая функция обеспечивается тем, что в одном индексе обобщаются (синтезируются) непосредственно несоизмеримые явления. Например, цены на разные товары или товары, абсолютно не сопоставимые между собой в натуральном выражении. Когда мы записываем

$$I_p = \frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0},$$

то благодаря использованию ценового соизмерителя можно агрегировать данные по различным товарам.

Вторая функция — аналитическая — следует из взаимосвязи индексов. Дело в том, что практически каждый индекс можно рассматривать как составляющую некой системы индексов, в которой его роль сводится к измерению одного из факторов общего изменения сложного явления и вклада этого фактора в совокупное изменение. Например, индекс цен можно рассматривать как показатель влияния изменения цен на выручку от продажи. Такая трактовка базируется на следующей связи признаков:

количество × цена = выручка (или затраты на покупку), т.е.

$$q \times p = w. \qquad (13.9)$$

Системе признаков соответствует система индексов (т.е. показателей их изменений). Исходя из этого можно записать:

$$I_q = I_{w(q)} = \frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0}; \qquad I_p = I_{w(p)} = \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0}. \qquad (13.10)$$

Обратите внимание: эта запись соответствует трактовке индекса как метода анализа. Когда мы указываем $I_{w(q)}$ или $I_{w(p)}$, то имеем в виду измерение общего изменения результативного явления (в данном случае w) за счет одного из факторов (q или p). Конечно, можно ограничиться записью I_q и I_p — ничего не изменится по существу.

При построении агрегатных индексов удобно пользоваться такими понятиями, как «индексируемый признак» и «признак-вес».

Индексируемый признак — это признак, изменение которого характеризует данный индекс. Например, в I_q — это q, в i_p — это p. Значение индексируемого признака изменяется: отчетное значение сопоставляется с базисным.

Признак-вес выполняет функцию веса по отношению к индексируемому признаку; его значение в данном индексе принимается неизменным, так как он не должен искажать оценку изменения индексируемого признака. В I_q признаком-весом является p, а в I_p — q.

Индексируемый признак можно назвать *фактором* изменения общего результата, а признак-вес — *характеристикой условий*, в которых оценивается это изменение.

Если индексы рассматриваются в системе, то должна обеспечиваться взаимосвязь между ними. Например, в соответствии с (13.9) должно выполняться равенство

$$I_q \cdot I_p = I_w. \qquad (13.11)$$

Обратимся к формуле (13.10). Каждый из индексов показывает, как изменился тот или иной фактор при неизменности прочих условий: и в формуле индекса I_q, и в формуле I_p веса закреплены на базисном уровне. Это обеспечивает сопоставимость оценок изменений факторов. Однако равенство (13.11) не обеспечивается или, как говорят, не обеспечивается увязка индексов в систему:

$$\frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0} \cdot \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0} \neq \frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_0 p_0}.$$

То же происходит, если все индексы будут построены с отчетными весами:

$$\frac{\sum_{(j)} q_1 p_1}{\sum_{(j)} q_0 p_1} \cdot \frac{\sum_{(j)} q_1 p_1}{\sum_{(j)} q_1 p_0} \neq \frac{\sum_{(j)} q_1 p_1}{\sum_{(j)} q_0 p_0}.$$

Только когда взаимосвязанные индексы строятся с весами разных периодов, увязка их в систему выполняется:

$$\frac{\sum_{(j)} q_1 p_1}{\sum_{(j)} q_0 p_1} \cdot \frac{\sum_{(j)} q_0 p_1}{\sum_{(j)} q_0 p_0} = \frac{\sum_{(j)} q_1 p_1}{\sum_{(j)} q_0 p_0} \qquad (13.12)$$

или

$$\frac{\sum_{(j)} q_1 p_0}{\sum_{(j)} q_0 p_0} \cdot \frac{\sum_{(j)} q_1 p_1}{\sum_{(j)} q_1 p_0} = \frac{\sum_{(j)} q_1 p_1}{\sum_{(j)} q_0 p_0}. \qquad (13.13)$$

Из этих двух вариантов отечественная статистика долгое время отдавала предпочтение второму. Соответственно существовало правило определения периода весов: индексы первичных признаков строятся на весах базисного периода, вторичных — на весах отчетного периода. Это правило признавало неравное значение признаков в системе: первичный признак выступает как основа формирования нового (отчетного) значения результативного признака w_1. Этим объясняется то, что индекс первичного признака (например, I_q) оценивает изменение данного признака при сохранении базисных условий (p_0), тогда как изменение вторичного признака оценивается уже в изменившихся условиях, когда первичный признак принял значение отчетного периода (q_1).

Пример. Рассмотрим, как влияет использование разных значений признака-веса на величину индекса (табл. 13.5).

1) с весами p_1 $I_q = \frac{1217,75}{1240,69} = 0,9815$, или 98,15%; с весами q_0:

$$I_p = \frac{1240,69}{1224,0} = 1,0136, \text{ или } 101,36\%;$$

$$I_w = \frac{1217,75}{1224,0} = 0,9949, \text{ или } 99,49\%; \; 0,9815 \cdot 1,0136 = 0,9949;$$

Таблица 13.5

Данные о продаже продуктов на городском рынке за месяц

Продукт	Цена, руб./кг		Продано, т		Выручка, тыс. руб.			
	май	июнь	май	июнь	май	июнь	условная	
	p_0	p_1	q_0	q_1	$w_0 = q_0 p_0$	$w_1 = q_1 p_1$	$q_1 p_0$	$q_0 p_1$
Говядина	125,5	130,7	3,0	2,98	376,50	389,49	373,99	392,10
Свинина	100,3	108,5	2,8	2,75	280,84	298,37	275,82	303,80
Картофель	12,7	13,4	10,2	10,80	129,54	144,72	137,16	136,68
Капуста	24,2	23,0	8,5	8,80	205,70	202,40	212,96	195,50
Яблоки	40,6	37,3	5,7	4,90	231,42	182,77	198,94	212,61
Итого	X	X	X	X	1224,0	1217,75	1198,87	1240,69

2) с весами p_0: $I_q = \frac{1198,87}{1224,0} = 0,9795$, или 97,95%; с весами q_1:

$$I_p = \frac{1217,75}{1198,87} = 1,0157, \text{ или } 101,57\%;$$

$$I_w = \frac{1217,75}{1224,0} \; 0,9949; \; 0,9795 \cdot 1,0157 = 0,9949, \text{ или } 99,49\%.$$

В обоих вариантах получены показатели снижения объема продажи и роста цен, но в первом случае объем продажи снизился на 1,85%, цены повысились на 1,36%, а во втором объем продажи снизился на 2,05% и цены повысились на 1,57%. Следуя статистической логике, можно сказать, что точечные оценки изменений в принципе невозможны: речь может идти лишь о поле или интервале оценок: для объема продажи — снижение от −1,57% до −1,85%; для цен — рост от 1,36% до 1,57%.

Однако в практическом использовании индексов стремятся получить однозначное решение тем или иным способом. Первый способ — получение средних оценок изменений: либо в форме индексов, построенных на средних весах:

$$I_p = \frac{\sum_{(j)} p_1 \frac{q_1 + q_0}{2}}{\sum_{(j)} p_0 \frac{q_1 + q_0}{2}},$$

либо через осреднение разновзвешенных индексов. При этом предпочтение отдается средней геометрической (индекс Фишера):

$$I_p = \sqrt{\frac{\sum q_1 p_1}{\sum q_1 p_0} \cdot \frac{\sum q_0 p_1}{\sum q_0 p_0}}. \qquad (13.14)$$

Второй способ основан на предпочтении какого-то одного варианта построения взаимосвязанных индексов. Как уже отмечалось, в отечественной статистике был принят второй способ. Но при этом возникала несопоставимость оценок изменений признаков. Поэтому делалась попытка построения всех взаимосвязанных индексов на весах одного периода — базисного:

$$I_q = \frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0} \cdot \quad I_p = \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0}. \qquad (13.15)$$

Понятно, что в этом случае не выполняется увязка индексов в систему:

$$I_q \cdot I_p \neq I_w.$$

Изолированная оценка изменения каждого фактора при неизменности другого приводит к недоучету эффекта совместного изменения факторов. Скажем, вы смотрите на движущееся изображение без звука или слушаете звуковое сопровождение без изображения, и в том, и в другом случае воздействие меньше, чем при соединении изображения и звука. Наглядно это можно показать с помощью особого вида плоскостной диаграммы, известной в отечественной статистике как «знак Варзара» (по имени российского статистика В. Е. Варзара (1851—1940), рис. 13.1.

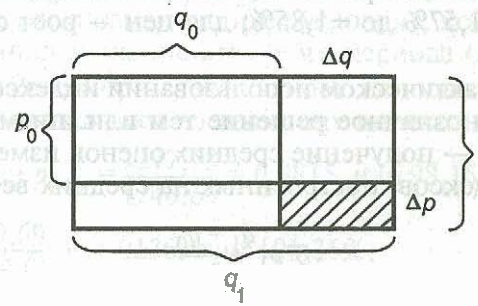


Рис. 13.1. Знак Варзара

Результативное явление представлено здесь в виде прямоугольника, площадь которого в базисном периоде $s_{q_0p_0}$, в отчетном — $s_{q_1p_1}$. Переход от базисного состояния к отчетному формируется за счет изменения фактора $q(s_{\Delta qp_0})$, изменения фактора $p(s_{\Delta pq_0})$ и совместного изменения обоих факторов $(s_{\Delta q\Delta p})$:

$$S_{q_1p_1} = s_{q_0p_0} + s_{\Delta qp_0} + s_{\Delta pq_0} + s_{\Delta q\Delta p}. \qquad (13.16)$$

В статистической науке выработано множество версий аналитического разложения: 1) выделение эффекта взаимодействия факторов в самостоятельный член; 2) присоединение его к какому-либо одному фактору (т.е. построение какого-либо из индексов на весах отчетного периода); 3) разделение эффекта взаимодействия факторов и присоединение к изменениям факторов — поровну либо пропорционально значениям индексов факторов, либо еще по какому-то принципу. Вы можете тоже попытаться предложить свое решение — актуальность проблемы сохраняется.

В. И. Борткевич (1868—1931) вывел формулу, объясняющую различие между индексами с разными весами:

$$\underbrace{\frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_1 p_0} : \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0}} = 1 + r_{i_q i_p} v_{i_q} v_{i_p}. \qquad (13.17)$$

Формула (13.17) позволяет измерить эффект совместного изменения факторов q и p

Точно так же можно выразить соотношение между индексами фактора q с разными весами. Из формулы (13.17) ясно, что индексы с отчетными и базисными весами будут равны, если выполняется хотя бы одно из условий: или корреляция между изменениями цен и объема продажи на отдельные товары отсутствует: $r_{i_p i_q} = 0$; или темпы изменения объемов то-

варов всех видов будут одинаковы: $v_{i_q} = 0$; или темпы изменения цен на все товары будут одинаковы: $v_{i_p} = 0$. Чем большая дистанция разделяет сравнимые периоды, тем сильнее проявляются все отмеченные факторы различий между индексами с разными весами.

Ничего не меняется, если результативный признак включает более двух факторов, т.е. в случае мультипликативной модели:

$$y = x_1 \cdot x_2 \cdot \ldots \cdot x_k.$$

Если придерживаться концепции неравноправия факторов и строить индексы с разными весами, то все зависит от принятой последовательности факторов в системе. Например, общие затраты на кожу для изготовления женских туфель можно представить как $w = qlp$, где q — количество пар туфель; l — средний расход кожи на одну пару; p — цена кожи. На первом месте стоит фактор q как первичный, с которого и начинаются все изменения. Тогда индексы будут иметь вид:

$$\frac{\sum q_1 l_0 p_0}{\sum q_0 l_0 p_0} \cdot \frac{\sum q_1 l_1 p_0}{\sum q_1 l_0 p_0} \cdot \frac{\sum q_1 l_1 p_1}{\sum q_1 l_1 p_0} = \frac{\sum q_1 l_1 p_1}{\sum q_0 l_0 p_0} = \frac{\sum w_1}{\sum w_0}. \quad (13.18)$$

Здесь применяется то же правило выбора весов, которое было сформулировано выше. Признаки, стоящие слева от индексируемого признака, трактуются по отношению к нему как первичные и закрепляются на отчетном уровне (они «уже» изменились), стоящие справа от него трактуются как вторичные и закрепляются на базисном уровне (они как бы «еще» не изменились). К этому добавляется условие содержательной интерпретации при последовательном объединении признаков слева направо. Скажем, произведение ql имеет экономический смысл — это расход кожи на весь объем производства туфель, при перестановке признаков q, p, l произведение qp экономического смысла не имеет. На таком подходе основан метод цепных подстановок, широко используемый в экономическом анализе.

Если же все индексы строятся на весах одного и того же (базисного) периода, то последовательность признаков не имеет значения. Система индексов будет иметь вид:

$$\frac{\sum q_1 l_0 p_0}{\sum q_0 l_0 p_0} \cdot \frac{\sum q_0 l_1 p_0}{\sum q_0 l_0 p_0} \cdot \frac{\sum q_0 l_0 p_1}{\sum q_0 l_0 p_0} \cdot I_{\substack{\text{совместных} \\ \text{изменений}}} = I_w; \quad (13.19)$$

$$I_{\substack{\text{совместных} \\ \text{изменений}}} = \frac{I_w}{I_q \cdot I_l \cdot I_p}.$$

И в случае многофакторной модели эффект совместных изменений можно либо сохранить в качестве самостоятельного члена разложения, либо распределить между изменениями факторов. Это зависит от поставленной задачи и от пристрастий исследователя.

Сравнение данных отчетного и базисного периодов неявно предполагает представление экономических процессов в виде дискретной последовательности периодов времени, что особенно проблематично при сравнении в длительном периоде. Экономические индексы для моментов непрерывного времени были предложены в 1928 г. французским статистиком Ф. Девизиа. Это привело к использованию в индексном анализе дифференциального исчисления. Данный подход до сих пор не вошел в статистическую практику, однако теоретически он более обоснован, нежели традиционные методы.

13.4. Свойства индексов

Как было показано, при построении индексов возникает много дискуссионных вопросов. Индексы считаются построенными правильно, если они удовлетворяют ряду тестов. Эти тесты были сформулированы американским статистиком И. Фишером (1867—1947). Основные тесты таковы.

1. Тест обратимости во времени. Индексы, исчисленные в «прямом» и «обратном» направлениях, должны быть взаимообратными числами. Например, если индекс показывает, что уровень цен в отчетном периоде по сравнению с базисным повысился в два раза, то он должен отражать, что в базисном периоде цены были вполовину ниже, чем в отчетном, т.е.

$$I_{ab} \cdot I_{ba} = 1, \qquad (13.20)$$

где a и b — сравниваемые периоды.

Очевидно, что наличие этого свойства желательно у любого индекса, ибо в таком случае сравнение между двумя состояниями не будет зависеть от того, какое из них принято за базу, особенно это важно при территориальных сравнениях.

2. Тест обратимости по факторам. Если поменять местами в индексе цен символы для цен и для количества, то мы должны получить индекс количества, который, будучи умножен на индекс цен, должен дать изменение общей стоимости товаров. Например, имеем:

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}.$$

Если теперь поменять местами p и q, то получим:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}.$$

Произведение этих индексов

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} \cdot \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

не равно индексу общей стоимости $\sum q_1 p_1 / \sum q_0 p_0$. Следовательно, индексы этого типа не отвечают тесту обратимости факторов.

Данному тесту отвечает средний геометрический индекс (13.14). По этой причине он был назван И. Фишером идеальным индексом.

3. Тест кружного испытания (циркулярность). Если построен некоторый индекс для года a при базисном годе b и для года b при базисном годе c, то из них можно получить индекс года a при базисном годе c. Тест кружного испытания требует, чтобы I_{ac}, основанный на промежуточных сравнениях, совпал с тем, какой мы получили бы при непосредственном сравнении $_{ac}$, т.е.

$$I_{ab} \cdot I_{bc} = I_{ac}. \qquad (13.21)$$

Это требование принято называть в статистике «цепным тестом».

В случае взвешенных индексов этот тест выполняется только для индексов с постоянными весами. Особенно трудно обеспечить выполнение данного теста при сравнении с отдаленной базой.

Легко сравнивать каждый из ряда лет с предыдущим, но нелегко сравнивать удаленные годы: произведение цепных сравнений (т.е. прилежащих годов) может отличаться от результатов непосредственного сравнения лет в начале и конце периода. Тут возникает много экономических проблем — и постоянство весов (проблема выбора неизменных цен при построении индексов объема производства), и выделение сравнимого круга элементов на протяжении всего периода (сравнимого круга товаров, видов продукции, труда и т.д.) при анализе изменений цен, заработной платы и т.п.

В этот же тест Фишер вводил условие *круговой сходимости*, которое гласит: если условия начального и конечного моментов времени совпадают по уровням цен и объемов товаров, то произведение индексов цен и объемов товаров за все подпериоды должно быть равно единице.

4. Соизмеримость. Численные значения индексов не должны зависеть от выбора единиц измерения объемов товаров и цен.

5. Пропорциональность. Согласно данному тесту если темпы роста всех цен (или объемов товаров) равны одному и тому же числу, то этому же числу должен быть равен индекс цен (или индекс объема).

6. Включение-исключение. Если к набору товаров, по которым вычисляются индексы, и объему товаров добавить еще один товар, темпы роста цены (или объема) которого совпадают с первоначальным индексом, то первоначальный индекс цен (или объема) не должен измениться.

Как видим, формулировка всех тестов основана на логике построения экономико-статистических показателей.

Тесты И. Фишера сыграли большую роль в развитии методологии экономических индексов.

13.5. Индексный анализ взвешенной средней. Индекс структуры

Индексы позволяют анализировать изменения не только агрегатов, но и средних величин. Предположим, изучается динамика средней цены товара на трех рынках города, расположенных в разных районах — центральном и двух периферийных — старой и новой застройки. Уровень цен в этих районах разный, соответственно на среднюю цену продажи на колхозных рынках влияют не только цены на каждом из них, но и доля каждого рынка в общем объеме продажи.

Формула средней цены:

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i q_i}{\sum_{i=1}^{n} q_i},$$

где p_i — цена товара на i-м рынке, $q_i / \sum_{(i)} q_i$ — структура продажи.

Изменение средней цены (как и любой взвешенной средней) выражается индексом:

$$I_{\bar{p}} = \frac{\bar{p}_1}{\bar{p}_0} = \frac{\sum_{(i)} q_i p_i}{\sum_{(i)} q_1} : \frac{\sum q_0 p_0}{\sum q_0}. \qquad (13.22)$$

Этот индекс получил название *индекс переменного состава*, так как отражает не только изменение осредняемого признака p, но и структуру совокупности $q_i / \sum_{(i)} q_i$. На основе индекса средней величины могут быть построены индексы самого осредняемого признака при постоянстве структуры совокупности и индекс структуры:

$$I_{\bar{p}} = \frac{\sum q_1 p_1}{\sum q_1} : \frac{\sum q_1 p_0}{\sum q_1} = \frac{\sum q_1 p_1}{\sum q_1 p_0}. \qquad (13.23)$$

Этот индекс получил название *индекс постоянного состава*.

Соответственно

$$I_{\text{структуры}} = \frac{\sum q_1 p_0}{\sum q_1} : \frac{\sum q_0 p_0}{\sum q_0}. \qquad (13.24)$$

Формулы индексов (13.23) и (13.24) основаны на общепринятом правиле, по которому структура совокупности как первичная характеристика при индексации цен закрепляется на уровне отчетного периода, а цены как вторичная характеристика при индексации структуры закрепляются на уровне базисного периода. Очевидно, что применение весов различных периодов и в этом случае обеспечивает выполнение равенства:

$$I_p \cdot I_{\text{структуры}} = I_{\bar{p}} \qquad \text{или} \qquad I_{\bar{p}(p)} \cdot I_{\bar{p}\left(\frac{q}{\sum q}\right)} = I_{\bar{p}}. \qquad (13.25)$$

Конечно, можно все индексы построить на весах базисного периода, и это будет правильнее с точки зрения оценки изменения каждого из факторов, но тогда равенство (13.25) будет нарушено.

Пример. Рассмотрим построение этих индексов. На трех рынках города продается картофель. Данные о продаже за день в зарегистрированных ценах приведены в табл. 13.6.

Таблица 13.6

Дневная продажа картофеля на колхозных рынках города

Рынок	Объем дневной продажи, кг		Цена, руб./кг		Изменение цены, %	Удельный вес рынка в объеме продажи, %		Выручка от продажи, руб.		
	август	сентябрь	август	сентябрь				август	сентябрь	условная
	q_0	q_1	p_0	p_1	i_p	d_0	d_1	q_0p_0	q_1p_1	q_1p_0
Центральный	160	170	17,0	16,5	97,1	38,1	30,9	2720	2805	2890
Старый	100	130	16,5	15,3	92,7	23,8	23,6	1650	1989	2145
Новый	160	250	18,0	17,2	95,6	38,1	45,5	2880	4300	4500
Итого	420	550	17,3	16,5		100,0	100,0	7250	9094	9535

Средняя цена картофеля в августе составила: $\bar{p}_0 = 17{,}26$ руб./кг, в сентябре: $\bar{p}_1 = 16{,}53$ руб./кг. Наибольшее снижение цен произошло на Старом рынке. Наибольшее увеличение объема продаж — на Новом рынке, в результате чего доля этого рынка в общей дневной реализации картофеля в сентябре стала составлять почти половину всего объема.

Индекс средней цены составил:

$I_{\bar{p}} = 16{,}53$ руб./кг : 17,26 руб./кг = 0,9577 = 100% = 95,77% (−4,23%).

Изменение самой цены в условиях структуры продажи, сложившейся в отчетном периоде, составило:

$$I_p = 16{,}53 : \frac{9535}{550} = 16{,}53 : 17{,}34 = 0{,}953\%,$$

т.е. среднее снижение цен на рынках было несколько большим, чем снижение средней цены (−4,7% против −4,23%). Индекс изменения цены был получен делением средней цены в отчетном периоде на среднюю условную цену, которая была бы при базисном уровне цен на рынках и отчетной структуре продаж. Этот же индекс можно было получить как отношение сумм выручки в отчетном периоде к условной выручке:

$$I_p = \frac{9094,\ \text{руб.}}{9535,\ \text{руб.}} = 0{,}953.$$

Индекс изменения цен — это индекс постоянного состава. Различие между этим индексом I_p и индексом средней цены (т.е. индексом переменного состава) $I_{\bar{p}}$ вызвано изменением структуры продаж:

$$I_{\text{структуры}} = 17{,}34 \text{ руб./кг}: 17{,}26 \text{ руб./кг} = 1{,}0046, \text{ или } 100{,}5\%.$$

За счет изменения структуры продажи средняя цена картофеля возросла почти на 0,5%, или на 8 коп./кг (17,34 − 17,26). Это связано с повышением удельного веса Нового рынка, на котором цены выше. Очевидно, что выполняется равенство: $0{,}953 \cdot 1{,}0056 = 0{,}9577$, или 95,77%.

Если использовать обозначение структуры продажи d, то индексы (13.22), (13.23), (13.24) будут иметь вид:

$$\frac{\sum p_1 d_1}{\sum p_0 d_0} = \frac{\sum p_1 d_1}{\sum p_0 d_1} \cdot \frac{\sum p_0 d_1}{\sum p_0 d_0}. \qquad (13.26)$$

Можно выразить и абсолютное изменение средней величины с учетом изменения факторов — самого осредняемого признака и структуры (т.е. признака-веса):

$$\Delta\bar{p} = \Delta\bar{p}(p) + \Delta\bar{p}(d). \qquad (13.27)$$

По данным табл. 13.6 средняя цена картофеля понизилась на 4,3 руб./кг: $\Delta\bar{p} = 16{,}53 - 17{,}26 = -0{,}73$ руб./кг; в том числе за счет самой цены: $\Delta\bar{p}(p) = 16{,}53 - 17{,}34 = -0{,}81$ руб./кг и за счет структурного фактора: $\Delta\bar{p}(d) = 17{,}34 - 17{,}26 = 0{,}08$ руб./кг.

И при относительном, и при абсолютном разложении эффект взаимодействия факторов — цены и структуры продажи — присоединился к оценке изменения цен. Если получить эту оценку в условиях базисного периода, то сравнение индексов

$$I_p = \frac{\sum p_1 d_1}{\sum p_0 d_1} \quad \text{и} \quad I_p = \frac{\sum p_1 d_0}{\sum p_0 d_0}$$

позволит выделить эффект совместного изменения факторов. По данным табл. 13.6 получаем:

$$I_p = \frac{\sum p_1 d_0}{\sum p_0 d_0} = \frac{1648}{1726} = 0{,}9548.$$

Этот результат мало отличается от того, который был получен в условиях структуры продажи отчетного периода ($I_p = 0{,}953$), так что эффект взаимодействия факторов оказался незначителен и направлен на повышение средней цены.

Влияние структурных сдвигов иногда приводит к неожиданным результатам: изменение себестоимости в целом по отрасли может оказаться большим, чем на отдельных предприятиях; или при выполнении производственной программы всеми предприятиями региона может оказаться, что регион в целом с программой не справился. Этот вопрос подробнее освещен в подразд. 13.7.

13.6. Построение индексов при обобщении данных по единицам совокупности и по элементам

Мы обсудили построение индексов при обобщении данных по многим товарам или элементам и при обобщении данных по единицам при наличии одного элемента (одного вида товара). В экономических расчетах приходится иметь дело с задачами построения индексов, объединяющих данные по единицам и по элементам.

Обозначим число элементов m, число единиц n. Обобщение данных при построении индексов можно подразделить на три уровня:

1) $n = 1$, $m > 1$ — индексный анализ проводится по одной единице (предприятию, магазину и т.д.) и группе элементов;

2) $n > 1$, $m = 1$ — индексный анализ по группе единиц и одному элементу (товару, виду продукции);

3) $n > 1$, $m > 1$ — индексный анализ по группе единиц и элементов.

Последний тип задач характерен для муниципального управления, аналитической работы в региональных, ведомственных и федеральных статистических службах.

Предположим, нам нужно изучить потребление продовольствия в районе города. Собраны данные об объеме покупок товаров и ценах на рынках и в магазинах (табл. 13.7).

По данным табл. 13.7 можно построить индексы объема продажи и це ля каждого вида торговли в отдельности, что соответствует нным типа 1; можно определить изменение объема прода и цен на каждый из товаров по всем видам торговли, что ответствует данным типа 2; наконец, можно получить индексы объема продажи и цен по всем видам торговли и всем товарам, что соответствует данным типа 3.

Проведем последовательно расчеты всех индексов, используя базисные веса в индексах объема продаж и отчетные — в индексах цен.

Пример. Изучается потребление продовольствия в районе города. Собраны данные об объеме покупок товаров и ценах на рынке и в магазинах (см. табл. 13.7).

По данным табл. 13.7 можно построить индексы объема продажи и цен для каждого вида торговли в отдельности, что соответствует данным типа 1; можно определить изменение

Таблица 13.7

Данные о продаже продовольственных товаров в районе города

Организация	Товар	Базисный период			Отчетный период			Расчетная выручка	
		т q_0	руб./кг p_0	тыс. руб. q_0p_0	т q_1	руб./кг p_1	тыс. руб. q_1p_1	тыс. руб. q_1p_0	тыс. руб. q_0p_1
Рынок	Говядина	2,2	87	191,4	1,5	120	180,0	130,5	264,0
	Свинина	1,3	90	117,0	1,5	135	202,5	135,0	175,5
	Картофель	4,5	10	45,0	4,3	12	51,6	43,0	54,0
	Яблоки	4,0	37	148,0	3,5	45	157,5	129,5	180,0
	Творог	1,0	100	100,0	1,2	110	132,0	120,0	110,0
Итого	По полному кругу	—	—	601,4	—	—	723,6	558,0	783,5
Магазин 1	Говядина	3,0	68	204,0	3,5	72	252,0	238,0	216,0
	Свинина	—	—	—	2,0	70	140,0	(132,4)	—
	Картофель	3,0	7	21,0	2,7	8	21,6	18,9	24,0
	Яблоки	2,5	25	62,5	2,1	27	56,7	52,5	67,5
	Творог	1,2	80	96,0	1,0	85	85,0	80,0	102,0
	Сахарный песок	4,0	15	60,0	4,5	19	85,5	67,5	76,0
Итого	По полному кругу	—	—	443,5	—	—	640,8	(589,3)	485,5
В том числе	По сопоставимому кругу				—	—	500,8	456,9	485,5

Продолжение

Организация	Товар	Базисный период			Отчетный период			Расчетная выручка	
		т q_0	руб./кг p_0	тыс. руб. q_0p_0	т q_1	руб./кг p_1	тыс. руб. q_1p_1	тыс. руб. q_1p_0	тыс. руб. q_0p_1
Магазин 2	Говядина	1,0	70	70,0	—	—	—	—	(70)
	Свинина	0,8	72	57,6	0,8	75	60,0	57,6	60
	Яблоки	1,5	30	45,0	1,0	37	37,0	30,0	55,5
	Сахарный песок	2,0	17	34,0	2,1	21	44,1	35,7	42
Итого	По полному кругу	—	—	206,6	—	—	141,1	123,3	(227,5)
В том числе	По сопоставимому кругу	—	—	136,6	—	—	141,1	123,3	157,5
		Q_0	$\bar{p}_0$	$Q_0\bar{p}_0$	Q_1	$\bar{p}_1$	$Q_1\bar{p}_1$	$Q_1\bar{p}_0$	$Q_0\bar{p}_1$
Всего по всем организациям	Говядина	6,2	75,1	465,4	5,0	86,4	432,0	375,5	535,7
	Свинина	2,1	83,1	174,6	4,3	93,6	402,5	357,3	196,6
	Картофель	7,5	8,8	66,0	7,0	10,4	73,2	61,6	78,0
	Яблоки	8,0	31,9	255,5	6,6	38,1	251,2	210,6	304,8
	Творог	2,2	89,1	196,0	2,2	98,6	217,0	196,0	216,9
	Сахарный песок	6,0	15,7	94,0	6,6	19,6	129,6	103,6	117,6
Итого		—	—	1251,5	—	—	1505,5	1304,6	1449,6

объема продажи и цен на каждый из товаров по всем видам торговли, что соответствует данным типа 2; наконец, можно получить индексы объема продажи и цен по всем видам торговли и всем товарам, что соответствует данным типа 3. Проведем последовательно расчеты всех индексов, используя базисные веса в индексах объема продажи и отчетные — в индексах цен.

Рынок.

$$I_q = \frac{558,0}{601,4} = 0,928 \cdot 100\% = 92,8\ \%\ (-7,2\%);$$

$$I_p = \frac{723,6}{558,0} = 1,297 \cdot 100\% = 129,7\%\ (+29,7\%);$$

$$I_w = \frac{723,6}{601,4} = 1,203 \cdot 100\% = 120,3\%\ (+20,3\%).$$

Магазин 1. В этом магазине ассортимент товаров изменялся: в базисном периоде свинины не было в продаже, в отчетном — она появилась. В этом случае изменение цен определяется по сопоставимому кругу товаров, т.е. по пяти товарам.

Обозначим сопоставимый круг товаров l, полный круг в базисном периоде — m_0, в отчетном — m_1. Тогда:

$$I_p = \frac{\sum\limits_{1}^{l} q_1 p_1}{\sum\limits_{1}^{l} q_1 p_0}.$$

По данным табл. 13.7

$$I_p = \frac{500,8}{456,9} = 1,096 \cdot 100\% = 109,6\%\ (+9,6\%).$$

Индекс объема продажи должен отразить изменение объема продажи тех товаров, которые были в базисном и продолжали продаваться в отчетном периоде, и, кроме того, изменение в объеме продажи в связи с появлением нового товара (несопоставимого). Так что в числителе индекса — сумма выручки по полному кругу отчетного периода, в знаменателе — по полному кругу базисного периода:

$$I_q = \frac{\sum_{1}^{m_1} q_1 p_0}{\sum_{1}^{m_0} q_0 p_0},$$

т.е. этот индекс должен включать данные по всем товарам: сопоставимым и несопоставимым.

В нашем примере

$$I_q = \frac{\sum_{1}^{6} q_1 p_0}{\sum_{1}^{5} q_0 p_0}.$$

Возникла проблема определения базисной цены для товара, который имелся только в отчетном периоде (т.е. для свинины).

Возможны по крайней мере три способа ее решения.

1. Использование для несопоставимых элементов (для новых товаров) цен отчетного периода, т.е. числитель I_q представляется как:

$$\left(\sum_{1}^{l} q_1 p_0 + \sum_{1}^{m_1 - 1} q_1 p_1\right),$$

где $(m_1 - 1)$ — несопоставимый круг в отчетном периоде.

По данным таблицы 13.7 получим:

$$I_q = \frac{456{,}9 + 140}{443{,}5} = 1{,}346 \cdot 100\% = 134{,}6\% \ (+34{,}6\%).$$

Расчет I_q этим методом нарушает увязку индексов в систему: $1{,}346 \cdot 1{,}096 = 1{,}475$ (1 + 47,5%), тогда как по данным табл. 13.7 индекс выручки составил:

$$I_w = I_{qp} = \frac{640{,}8}{443{,}5} = 1{,}445 \cdot 100\% = 144{,}5\%;$$

$$144{,}5\% \neq 147{,}5\%.$$

Однако выполняется увязка абсолютных изменений:

$$\Delta_w = \Delta_{qp} = \Delta_q p_0 + \Delta_p q_0 = 153{,}4 + 43{,}9 =$$
$$= (596{,}9 - 443{,}5) + (500{,}8 - 456{,}9) = 197{,}3 \text{ тыс. руб.}$$

Если сравним непосредственно суммы выручки отчетного и базисного периодов, то получим ту же величину:

$$\Delta_w = 640{,}3 - 443{,}5 = 197{,}3 \text{ тыс. руб.} -$$

прирост выручки в магазине 1.

2. Использование условных значений базисных цен, которые определяются расчетным путем. Логично предположить, что если бы свинина была в базисном периоде, то цены на нее повысились бы примерно так же, как и на остальные товары. Это предположение можно записать в виде равенства:

$$\frac{p_1^c}{p_{0(усл)}^c} = I_p, \text{ отсюда: } p_{0(усл)}^c = \frac{p_1^c}{I_p}.$$

По данным примера (табл. 13.7) условная цена свинины в магазине 1 была бы равна:

$$p_0^c = \frac{70}{1{,}096} = 63{,}9 \text{ руб/кг.}$$

3. Использование базисных цен других единиц или средних цен по совокупности. Например, в расчете индекса объема продажи в магазине 1 будем использовать базисную цену на свинину в магазине 2. В этом случае

$$I_q = \frac{456{,}9 + 2 \cdot 72}{443{,}5} = 1{,}355 \cdot 100\% = 135{,}5\%.$$

Или же используем в расчете I_q средние цены на свинину в базисном периоде. Поскольку значительный объем продаж свинины осуществляется на рынке, где цены выше, то средняя цена выше цены на свинину в магазине 2: $\bar{p}_0^c = 83{,}1$. Использование средней базисной цены приведет к еще более высокому значению индекса продажи в магазине 1:

$$I_q = \frac{456{,}9 + 2 \cdot 83{,}1}{443{,}5} = \frac{623{,}1}{443{,}5} = 1{,}495 \cdot 100\% = 140{,}5\%.$$

Это значение индекса превышает все предыдущие. Таким образом, на вопрос, как изменился объем продажи в магазине 1, трудно ответить однозначно: оценки варьируют от 131,8 до 140,5%. По-видимому, более реальны оценки роста объема продажи на 31,8% или на 34,6%.

Магазин 2. Здесь тоже ассортимент менялся: в отчетном периоде не торговали говядиной. Но это не вызывает трудностей в построении индексов, поскольку изменение объема продаж обусловлено как изменением продаж сопоставимых товаров, так и отсутствием продажи несопоставимого товара.

Первым рассчитывается индекс цен по сопоставимому кругу товаров:

$$I_p = \frac{\sum_{1}^{3} q_1 p_1}{\sum_{1}^{3} q_1 p_0} = \frac{141,1}{123,3} = 1,144 \cdot 100\% = 114,4\% -$$

в среднем по трем товарам цены повысились на 14,4%, т.е. больше, чем в магазине 1. Затем вычисляется индекс объема продажи без всяких условностей по тем товарам, что реально продавались; в отчетном периоде их было три, в базисном — четыре.

$$I_q = \frac{\sum_{1}^{3} q_1 p_0}{\sum_{1}^{4} q_0 p_0} = \frac{123,3}{206,6} = 0,597 \cdot 100\% = 59,7\%.$$

Если нужно рассчитать $q_0 p_1$ для несопоставимого товара, то можно воспользоваться одним из рассмотренных способов.

После этого вычислим индекс выручки:

$$I_{qp} = \frac{141,1}{206,6} = 0,683\ (68,3\%) \text{ или } 1,144 \cdot 0,597 = 0,683.$$

Для того чтобы получить результаты по всем видам торговли, данные обобщаются определенным образом. При этом возможны два подхода.

Первый подход основан на суммировании данных по видам торговли (или отдельным предприятиям). Этот метод базируется на данных отдельных хозяйственных единиц и поэтому называется *заводским методом.*

$$I_q = \frac{\sum_{1}^{n}\sum_{1}^{m_1} q_1 p_0}{\sum_{1}^{n}\sum_{1}^{m_1} q_0 p_0}; \qquad (13.28)$$

$$I_p = \frac{\sum_{1}^{n}\sum_{1}^{l} q_1 p_1}{\sum_{1}^{n}\sum_{1}^{l} q_1 p_0}, \qquad (13.29)$$

где n — число единиц совокупности;
m — число элементов, всего;
l — число сопоставимых элементов.

В этом случае один и тот же товар или вид продукции взвешивается по разным ценам в зависимости от того, где он учитывается.

Второй подход основан на обобщении данных по отдельным товарам независимо от места реализации. Для каждого товара рассчитываются сводные показатели количества и цены:

$$\sum_{1}^{n} q_1 = Q_1; \qquad \sum_{1}^{n} q_0 = Q_0;$$

$$\bar{p}_{1i} = \frac{\sum_{1}^{n} q_1 p_1}{\sum_{1}^{n} q_1}; \qquad \bar{p}_{0i} = \frac{\sum_{1}^{n} q_0 p_0}{\sum_{1}^{n} q_0}.$$

Затем данные обобщаются по всем товарам, при этом каждый из них взвешивается по средней цене для данного товара. Этот метод основан на обобщении с позиции совокупности, а не отдельных единиц и потому получил название *отраслевой метод*:

$$I_Q = \frac{\sum_{1}^{m_1} Q_1 \bar{p}_0}{\sum_{1}^{m_0} Q_0 \bar{p}_0}; \qquad (13.30)$$

$$I_p = \frac{\sum_{1}^{l} Q_1 \bar{p}_1}{\sum_{1}^{l} Q_1 \bar{p}_0}. \qquad (13.31)$$

При этом сопоставимость элементов определяется исходя из условий совокупности в целом.

В нашем примере (табл. 13.7) были товары, не сопоставимые с позиций отдельных торговых организаций, но в целом для торговли все они сопоставимы.

Вычислим индексы тем и другим методом:

1) по данным отдельных видов торговли

$$I_p = \frac{723{,}6 + 500{,}8 + 141{,}1}{558{,}0 + 456{,}9 + 123{,}3} = \frac{1365{,}5}{1138{,}2} = 1{,}1997 \cdot 100\% = 120\% \ (+20\%);$$

$$I_q = \frac{558{,}0 + 589{,}3 + 123{,}3}{601{,}4 + 443{,}5 + 206{,}6} = \frac{1270{,}6}{1251{,}5} = 1{,}015 \cdot 100\% = 101{,}5\%.$$

В целом объем продажи вырос в отчетном периоде на 1,5%, а цены повысились на 20%. Выручка от продажи возросла на 21,8%:

$$I_w = I_q \cdot I_p = 1{,}015 \cdot 1{,}1997 = 1{,}218 \cdot 100\% = 121{,}8\%;$$

2) по данным всех видов торговли, вместе взятых

$$I_Q = \frac{1304{,}6}{1251{,}5} = 1{,}042 \cdot 100\% = 104{,}2\%;$$

$$I_p = \frac{1505{,}5}{1304{,}6} = 1{,}154 \cdot 100\% = 115{,}4\%.$$

Получилось, что средние цены повысились в меньшей степени, чем в отдельных видах торговли: 15,4% против 20%. Это соотношение отражает влияние структурного фактора, изменение удельного веса продажи того или иного товара разными организациями. В частности, большое значение имела

продажа свинины в отчетном периоде не только на рынке, но и в магазине 1, где цены ниже.

Соотношение индекса средних цен и индекса цен без учета структурного фактора дает оценку структурных сдвигов:

$$I_{\text{структуры}} = I_{\bar{p}} : I_p = 1{,}154 : 1{,}20 = 0{,}962 \cdot 100\% = 96{,}2\%,$$

т.е. за счет изменения роли разных видов в общем объеме реализации средние цены снизились на 3,8%.

Мы получили индекс структуры исходя из взаимосвязи индексов. Можно рассчитать его значение непосредственно по формуле

$$I_{\text{структуры}} = \frac{\sum_{1}^{n}\sum_{1}^{m} q_1 p_0}{\sum_{1}^{m} Q_1 \bar{p}_0} = \frac{558{,}0 + 456{,}9 + 123{,}3}{1304{,}6} = \frac{1270{,}6}{1304{,}6} =$$

$$= 0{,}962 \cdot 100\% = 96{,}2\%.$$

Если стоит задача измерить влияние объема продаж на величину выручки от продажи, т.е. найти $I_{w(q)}$, то влияние измеряет индекс, найденный по заводскому методу. Его величина отражает одновременно и изменение объема продажи во всех видах торговли, и изменение структуры продажи.

Алгебраически это можно представить в виде равенства:

$$I_q = I_Q \cdot \underbrace{I_{\bar{p}\left(\frac{q}{\sum q}\right)}}; \qquad (13.32)$$

индекс структуры

$$\frac{\sum_{1}^{n}\sum_{1}^{m} q_1 p_0}{\sum_{1}^{n}\sum_{1}^{m} q_0 p_0} = \frac{\sum_{1}^{m} Q_1 \bar{p}_0}{\sum_{1}^{m} Q_0 \bar{p}_0} \cdot \frac{\sum_{1}^{n}\sum_{1}^{m} q_1 \bar{p}_0}{\sum_{1}^{m} Q_1 \bar{p}_0}, \qquad (13.33)$$

где

$$\sum_{1}^{n}\sum_{1}^{m} q_0 p_0 = \sum_{1}^{m} Q_0 \bar{p}_0 = \sum_{1}^{m} \frac{\sum_{1}^{n} q_0 p_0}{\sum_{1}^{n} q_0} \cdot \sum_{1}^{n} q_0 .$$

По данным табл. 13.7

$$I_Q = 1{,}042,\ I_{p\left(\frac{q}{\sum q}\right)} = 0{,}962;$$

тогда:

$$I_q = 1{,}042 \cdot 0{,}962 = 1{,}015 \cdot 100\% = 101{,}5\%.$$

Мы рассмотрели систему индексов, в которой использовались разновзвешенные индексы: индексы объемного фактора (количества) — с базисными весами, индексы качественного фактора (цен) — с отчетными весами. Плюсы и минусы такого метода построения индексов уже обсуждались.

Если необходимо все индексы построить на базисных весах, то в системе индексов появляется индекс совместных изменений. С экономической точки зрения его часто называют индексом смещения ассортимента, так как он показывает изменение в реализации доли товаров с разным типом снижения цен:

$$\frac{\sum_1^m Q_1\bar{p}_1}{\sum_1^m Q_1\bar{p}_0} = \frac{\sum_1^m Q_1\bar{p}_0}{\sum_1^m Q_0\bar{p}_0} \cdot \frac{\sum_1^m Q_0\bar{p}_1}{\sum_1^m Q_0\bar{p}_0} \left(\frac{\sum_1^m Q_1\bar{p}_1}{\sum_1^m Q_0\bar{p}_0} : \frac{\sum_1^m Q_0\bar{p}_1}{\sum_1^m Q_0\bar{p}_0} \right). \quad (13.34)$$

Можно оценить эффект совместного изменения признаков q и p в системе индексов, построенных по заводскому методу:

$$\frac{\sum\sum q_1p_1}{\sum\sum q_0p_0} = \frac{\sum\sum q_1p_0}{\sum\sum q_0p_0} \cdot \frac{\sum\sum q_0p_1}{\sum\sum q_0p_0} \left(\frac{\sum\sum q_1p_1}{\sum\sum q_1p_0} : \frac{\sum\sum q_0p_1}{\sum\sum q_0p_0} \right). \quad (13.35)$$

Вполне возможны различия в значениях индексов совместных изменений, полученных по формуле (13.34) или (13.35). Это различие может возникнуть из-за разного охвата элементов: в первом случае сопоставимость определяется с общеотраслевых позиций, во втором — с позиций отдельного предприятия.

Итак, вы получили представление о способах построения индексов при обобщении данных и по многим товарам, видам продукции, и по магазинам, рынкам, предприятиям. Какой способ выбрать в каждом конкретном случае, вам часто придется решать самим, ведь далеко не всегда имеется инструкция по проведению расчетов.

13.7. Границы и условия применения индексного метода

Каждый метод ориентирован на особые представления изучаемого объекта, на особую его модель. Индексный метод предполагает, что связь между признаками является жесткодетерминированной, которая проявляется как в каждом отдельном случае (для отдельного товара, вида продукции, предприятия и т.д.), так и в совокупности. Связь, изучаемая с помощью индексов, выражается в виде уравнения связи:

либо мультипликативного

$$y = x_1x_2\ldots x_k,$$

либо аддитивного

$$y = x_1 + x_2 + \ldots + x_k.$$

Вид функции, число переменных факторов (сомножителей или слагаемых) определяются нашими представлениями о логике изучаемой связи. Многофакторная мультипликативная модель строится путем последовательного расчленения одного из факторов на составляющие.

Например, можно записать следующее уравнение связи:

Объем произведенной продукции = Число отработанных человеко-часов · Средняя часовая выработка.

Эту модель можно детализировать. Она будет включать не два, а пять факторов:

Объем произведенной продукции = Средняя списочная численность работников · Доля рабочих в среднесписочной численности работников · Среднее число дней работы · Средняя продолжительность рабочего дня · Средняя часовая выработка.

Если мультипликативная модель имеет в качестве результативного первичный признак, то она называется полной[1]. Примером такой модели является вышеуказанная модель. Разделив обе части равенства на первый фактор, получим *неполную* модель среднечасовой выработки работника.

Представление связи как жесткодетерминированной является условным, так как связи социально-экономических явлений носят стохастический характер.

Если представить мультипликативную модель как двухфакторную, т.е. $y = x_1 x_2$, то в целом по совокупности уравнение имеет вид: $y = ax$. Коэффициент a является коэффициентом связи между y и x. Он передает *прямое* влияние фактора x на результат y. Для нашего примера величина отработанных человеко-часов передает влияние среднечасовой выработки на объем продукции. Однако выработка влияет на результат не только непосредственно, но и через другие факторы: уровень выработки может определять численность рабочих, их долю в списочном составе, фактическую продолжительность рабочего дня. В корреляционном анализе, измеряя корреляцию между результатом и фактором, мы получаем полную меру корреляции независимо от того, как реализуется связь — непосредственно или опосредованно. В индексном анализе мы измеряем только прямое влияние изменения фактора на изменение результата.

При построении уравнения связи иногда допускаются отступления от логики ради обеспечения увязки признаков, получения жесткодетерминированного выражения связи. Поэтому можно встретить уравнения связи, в которых не все составляющие элементы экономически обоснованы, нередки случаи появления среди факторов обратных величин.

Пример. Рассмотрим недостаточно обоснованное уравнение связи [1]:

$$\frac{\text{Балансовая прибыль}}{\text{Производственные фонды}} = \frac{\text{Балансовая прибыль}}{\text{Прибыль от реализации}} \cdot \frac{\text{Прибыль от реализации}}{\text{Стоимость реализованной продукции}} \cdot \frac{\text{Стоимость реализованной продукции}}{\text{Основные фонды}} \cdot \frac{\text{Основные фонды}}{\text{Производственные фонды}}.$$

[1] *Адамов В. Е.* Факторный индексный анализ. Методология и проблемы. — М.: Статистика, 1977. — С. 101.

Трудно представить, чтобы рост доли основных фондов вызывал рост балансовой прибыли.

При мультипликативной связи индексов относительные выражения приростов факторов связаны аддитивно. Например,

$$\frac{\sum q_1 z_1}{\sum q_0 z_0} = \frac{\sum q_1 z_0}{\sum q_0 z_0} \cdot \frac{\sum q_1 z_1}{\sum q_1 z_0},$$

однако

$$\frac{\Delta q x_0}{\sum q_0 z_0} = \frac{\sum q_1 z_0 - \sum q_0 z_0}{\sum q_0 z_0} + \frac{\Delta z q_1}{\sum q_0 z_0} = \frac{\sum q_1 z_1 - \sum q_1 z_0}{\sum q_0 z_0} = \frac{\sum q_1 z_1 - \sum q_0 z_0}{\sum q_0 z_0}.$$

Чем больше различаются индексы отдельных факторов, тем больше отличается сумма относительных приростов от темпа прироста результата. Например,

$$I_{\text{производ. труда}} \cdot I_{\text{затрат времени}} = I_{\text{объема продукции}}.$$

Если известно, что значение первого индекса 1,2, второго — 1,05, то

$$I_{\text{объема продукции}} = 1{,}26, \text{ тогда как } \frac{20\%}{100} + \frac{5\%}{100} = 25\%,$$

т.е. переход от темпов роста к темпам прироста приводит к определенным трудностям в интерпретации количественного влияния факторов на результат.

Наконец, решение вопроса об изменении эффекта отдельных факторов и их совместного изменения всегда условно.

Все предыдущее изложение было ориентировано на мультипликативную модель. При аддитивной связи признаков индексный анализ проводится по следующей формуле:

$$I_y = \frac{y_1}{y_0} = \frac{x_1 + z_1}{x_0 + z_0} = \frac{x_1}{x_0} \cdot \frac{x_0}{x_0 + z_0} + \frac{z_1}{z_0} \cdot \frac{z_0}{x_0 + z_0} = I_{x_{1/0}} d_{x_0} + I_{z_{1/0}} d_{z_0},$$

т.е. общее изменение результата зависит от изменения каждого фактора и его доли в базисной величине результата. Приведем пример (табл. 13.8).

Общее изменение численности работников может быть представлено как результат изменения численности занятых

Таблица 13.8

Численность работников на заводе

Период	Всего, чел.	В том числе	
		заняты физическим трудом	заняты умственным трудом
Базисный	1000	700	300
Отчетный	800	640	160

умственным и физическим трудом и их доли в общей численности работников:

$$\frac{640}{700} \cdot 0{,}7 + \frac{160}{300} \cdot 0{,}3 = 0{,}914 \cdot 0{,}7 + 0{,}533 \cdot 0{,}3 = 0{,}7997.$$

Этот результат отличается от 0,8 (800 : 1000) только за счет округления в расчетах.

Одним из сложнейших вопросов индексного анализа является оценка структурных сдвигов. Этот фактор может приводить к парадоксальным результатам в индексах.

Пример. Возьмем условные данные о работе трех химических предприятий одного района (табл. 13.9).

Таблица 13.9

Показатели работы химических предприятий района

Предприятие	Отчетный квартал		Прошлый квартал	
	товарная продукция в сопоставимых ценах, тыс. руб.	средняя списочная численность работающих, чел.	товарная продукция в сопоставимых ценах, тыс. руб.	средняя списочная численность работающих, чел.
Завод по производству минеральных удобрений	40 320	3200	14 400	1200
Фабрика искусственного меха	14 882	700	12 150	600
Завод пластиков	10 080	200	9600	200
В целом по химическим предприятиям	65 282	4100	36 150	2000

На каждом из этих предприятий рост объема производства сопровождался повышением производительности труда. Средняя выработка на одного работника по предприятиям составила в отчетном квартале соответственно, тыс. руб.: 12,6; 21,26 и 50,40. В прошлом квартале средняя выработка составляла, тыс. руб.: 12; 20,25 и 48. Сравнение этих данных показывает, что выработка росла равномерно на всех предприятиях: на первом заводе — 12,6 : 12 = 1,05, или 105%; на втором — 21,26 : 20,25 = 1,0498, или 104,98%; на третьем — 50,4 : 48 = = 1,05, или 105%. Если рассчитать среднюю выработку по всем трем заводам и определить ее динамику, то результат покажется невероятным. В отчетном квартале средняя выработка в целом составила 15,92 тыс. руб. (65 282 тыс. руб. : 4100 чел.), а в прошлом квартале — 18,075 тыс. руб. (36 150 тыс. руб. : 2000 чел.), т.е. средняя выработка по отрасли снизилась на 12% (15,92 тыс. руб. : 18,075 тыс. руб. = 0,88 · 100% = 88%).

Этот результат объясняется тем, что динамика среднеотраслевой выработки учитывает не только, какой была динамика выработки на отдельных предприятиях, но и как изменилось распределение работников между ними. Ведь уровень средней выработки на одного работника на отдельных предприятиях различается достаточно сильно: максимален он на заводе пластиков, минимален — на заводе минеральных удобрений. Именно на этом заводе численность работников возросла почти в 3 раза. Доля этого завода в численности работающих составляла 60% в прошлом квартале и 78,1% — в отчетном квартале. Отсюда и совокупный результат.

Можно измерить общее изменение выработки без учета изменения соотношений между предприятиями: если сравнить общий объем товарной продукции в сопоставимых ценах в отчетном квартале с тем объемом, который был бы получен, если бы выработка на каждом заводе оставалась прежней. Величина такой «условной» товарной продукции составит: 12 тыс. руб. × × 3200 чел. + 20,25 тыс. руб. · 700 чел. + 50,4 тыс. руб.× × 200 чел. = 62 175 тыс. руб. Суммарные показатели товарной продукции 65 282 тыс. руб. и 62 175 тыс. руб. различаются только за счет выработки, значит, их сравнение покажет динамику средней выработки по всем трем предприятиям без учета динамики численности работников. Действительно, получаем, что в целом рост выработки составил +5% (65 282 тыс. руб. :

: 62 175 тыс. руб. = 1,05, или 105%). Вот теперь нет никакого противоречия между результатами работы отдельных предприятий и отрасли. Но чтобы разобраться в этом, нужно знать, какой методикой пользовался статистик, как он получил те или иные результаты.

13.8. Комплексное использование индексного и регрессионного методов анализа

Применение индексного анализа часто оказывается недостаточным прежде всего из-за того, что уравнение связи как жесткодетерминированная функция может быть построено лишь для «ближайшего» круга факторов, тех, которые непосредственно составляют результат. Такие факторы могут оказаться недостаточными для объяснения его динамики. Эта особенность анализа связи на основе жесткодетерминированного выражения результата очевидна, например, при постатейном анализе себестоимости продукции. Вроде бы такой анализ обеспечивает точность показателей связи. Так, если изменятся норма расхода того или иного материала и заготовительные расходы на него, можно точно указать, на какую величину снизится (повысится) себестоимость продукции данного вида. Вместе с тем «функциональный» анализ себестоимости продукции вскрывает лишь непосредственное различие себестоимости из-за различий величин, прямо входящих в ее расчет, но не вскрывает причин самих этих различий. Можно установить, насколько на предприятиях, производящих однородную продукцию, различаются нормы расхода сырья, сдельные расценки и т.п. Но само по себе выяснение этих факторов еще ничего не говорит об их причинах, которые зависят от уровня технического оснащения предприятия, квалификации его работников, организации производства и т.п. Эти факторы воздействуют на величину себестоимости не непосредственно, а через величины, прямо учитываемые в ее расчетах: через нормы расхода материалов, расценки и ставки заработной платы, суммы амортизации и другие виды производственных затрат. В отличие от ближайших факторов такие факторы принадлежат к другому, так сказать глубинному уровню изучаемости структуры.

Далеко не всегда можно выявить механизмы связи между глубинными причинами и результатом в силу их большей отдаленности, многоплановости влияния. Не всегда можно включить их в жесткодетерминированное уравнение связи путем последовательного развертывания признаков. Это приводит к комплексному использованию методов, основанных на жесткой детерминации признаков, и методов, не ориентированных на такой характер связей.

Понять в полной мере задачи интеграции разных методов статистического изучения связей можно с помощью ***графа связей***. Граф связей учитывает непосредственные, т.е. причинные связи, которые предполагают изменение x_i при изменении влияющего на него x_j при постоянстве всех прочих факторов. Асимметричность причинных связей отражается в направленности дуг графа (*дуга* — соединение вершин графа, т.е. точек, соответствующих элементам структуры).

Разобраться в системе связей можно только тогда, когда граф связей будет включать не только факторы — признаки данной единицы совокупности, непосредственно определяемые в процессе ее функционирования (*эндогенные*), но и факторы, не зависящие от нее, но влияющие на изучаемый результат (*экзогенные*). Если первые образуют систему признаков и могут находиться в жесткодетерминированной связи с изучаемой результативной переменной вследствие устойчивости связи в рамках единицы совокупности, то вторые не являются признаками изучаемой единицы, потому их связь с результатом неустойчива, стохастична. Как правило, действие экзогенных факторов опосредовано эндогенными переменными, формирующими результат. Потребность сочетания разных уровней анализа — «вышележащего», на котором могут иметь место жесткодетерминированные связи, и «нижележащего», на котором они отсутствуют, вызывает интеграцию разных методов анализа. Так, изучая, почему произведен тот или иной объем валовой продукции, весьма важно не останавливаться на анализе уравнения связи, подобном приведенному в подразд. 13.7, включающем признаки, определяемые на уровне предприятия, а перейти на другой уровень анализа. Выявить, например, чем обусловлена та или иная величина среднечасовой выработки рабочих. Для этого необходимо перейти к совокупности рабочих и их признакам (уровню ква-

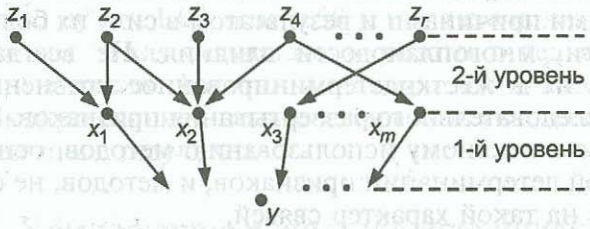


Рис. 13.2. Граф связей:
1-й уровень — жестко детерминированные связи; 2-й уровень — стохастические связи

лификации, стажу, умению организовывать процесс труда и т.д.).

На рис. 13.2 изображен гипотетический граф связей, в котором элементы «высшего» уровня структуры обозначены как x_j ($j = 1, 2, ..., m$). Это факторы, находящиеся в жесткой связи с результатом, y; z_l — глубинные факторы, принадлежащие другому уровню изучаемой структуры связей ($l = 1, 2, ..., z$). Эти факторы находятся в стохастической связи с x_j и y.

Рис. 13.2 показывает, что, используя индексы (или другой метод анализа, основанный на жесткодетерминированных связях), мы ограничиваемся только одним уровнем структуры связей, включающим отношения между y и x_j, и не затрагиваем связи между z_l и x_j. Используя лишь этот путь анализа, мы можем не выяснить причины изменения результата. Кроме того, в анализе только жесткодетерминированных связей [$x_j \to y$] каждый из x_j выступает как независимая величина, тогда как они могут быть связаны как непосредственно, так и через общие детерминирующие факторы. Эта связь является стохастической и может быть измерена с помощью соответствующих методов.

Методика комплексного использования индексного и регрессионного анализа такова. Определяется жесткодетерминированное уравнение связей: $y = f(x_1, ..., x_m)$, на основе графа связей строится уравнение регрессии для каждой компоненты x_j:

$$\hat{x}_j = a_0 + a_1 z_1 + ... + a_2 z_2,$$

где z_1 — так называемые глубинные причины.

Оценив значимость параметров отдельных регрессий, можно установить круг причин для каждого из x_j, общий круг причин для x_i и x_j ($i \neq j$). Используя полученное на основе регрессии значение $\hat{x}_j$, мы получаем возможность измерить влияние каждого из учтенных в регрессии факторов на y. Таким образом, в анализе участвуют функционально и нефункционально связанные факторы.

Остановимся подробнее на методике комплексного использования методов. Рассмотрим простейший случай. Пусть изучаемый результативный признак может быть представлен как жесткодетерминированная двухфакторная мультипликативная функция $y = xw$ (несмотря на то, что оба фактора x и w принадлежат к одному и тому же уровню изучаемой структуры, мы обозначили их по-разному, чтобы облегчить изложение методических вопросов). Пусть x — первичный (объемный) признак, w — вторичный (так называемый количественный) признак. Тогда система аналитических индексов имеет вид:

$$\frac{\sum x_1 w_1}{\sum x_0 w_0} \cdot \frac{\sum x_1 w_1}{\sum x_1 w_0} = \frac{\sum x_1 w_1}{\sum x_0 w_1} = \frac{\sum y_1}{\sum y_0}$$

или

$$I_{y(x)} \cdot I_{y(w)} = I_y.$$

На следующем этапе анализа перейдем на другой уровень структуры связей. Введем различные обозначения для факторов, влияющих на x и на w:

$$x = f(u_1, u_2, ..., m), \quad w = f/v_1, v_2, ...$$

Заметим, что в принципе, как уже отмечалось, круг факторов для x и w может частично совпадать. В случае непосредственной связи между x и w та из переменных, которая является независимой, может включаться в регрессию другой (зависимой) переменной. Предположим, что круг объясняющих переменных для x и w остался неизменным в отчетном периоде по сравнению с базисным. Принимая регрессии линейными, имеем по две регрессии для x и w, описывающих базисное и отчетное состояние x и w.

базисного периода:

$$\hat{x}_0 = a_{00} + a_{01} u_{01} + ... + a_{0m} u_{0m}$$

$$\hat{w}_0 = b_{00} + b_{01} \cdot v_{01} + \ldots + b_{0r} \cdot v_{0r}.$$

Для отчетного периода:

$$\hat{x}_1 = a_{10} + a_{11} \cdot u_{11} + \ldots + a_{1m} \cdot u_{1m};$$

$$\hat{w}_1 = b_{10} + b_{11} \cdot v_{11} + \ldots + b_{1r} \cdot v_{1r},$$

где первый подстрочный значок в каждой регрессии обозначает период, к которому она относится, второй — номер параметра или переменной соответственно.

Введем в индекс $I_{y(x)}$ расчетное значение $\hat{x}$, получим следующую систему индексов:

$$I_{y(x)} = \frac{\sum x_1 w_0}{\sum x_0 w_0} = \frac{\sum x_1 w_0}{\sum \hat{x}_1 w_0} \cdot \frac{\sum \hat{x}_1 w_0}{\sum \hat{x}_0 w_0} \cdot \frac{\sum \hat{x}_0 w_0}{\sum x_0 w_0}. \qquad (13.36)$$

Первый и последний индексы этой системы (13.36) измеряют влияние факторов, не учтенных в регрессии $x = \hat{f}(u_1, \ldots, u_m)$. Сравнение этих индексов позволяет установить, регрессия какого периода точнее описывает фактические данные. Если регрессии построены правильно, то расхождения фактических и расчетных значений x и для базисного, и для отчетного периодов будут незначительны, и оба индекса будут близки к единице.

Центральная роль принадлежит второму индексу системы — он измеряет влияние на y изменений в расчетных значениях $\hat{x}$. Расхождение между $\sum \hat{x}_1 w_0$ и $\sum \hat{x}_0 w_0$ может возникнуть как вследствие изменений значений переменных $u_1, \ldots, u_m$, так и в результате изменения силы их влияния на x — коэффициентов регрессии $a_{11}, a_{12}, \ldots, a_{1m}$ по сравнению с $a_{01}, a_{02}, \ldots, a_{0m}$. Раздельную оценку влияния на y глубинных факторов u и силу их воздействия a можно получить на основе специальной системы индексов. При этом первичным рекомендуется считать значение переменной, а вторичным — коэффициент регрессии[1].

Получим:

[1] *Юзбашев М., Рудакова Р.* Регрессионные модели и индексы в анализе сельскохозяйственных предприятий // Вестник статистики. — 1976. — № 5. — С. 56—66.

а) систему индексов, измеряющих влияние на y изменений значений переменных u:

$$I_{y\{\hat{x}(u_1)\}} = \frac{\sum (a_{00} + a_{01} \cdot u_{11} + \ldots + a_{0m} \cdot u_{01m}) W_0}{\sum (a_{00} + a_{01} \cdot u_{01} + \ldots + a_{0m} \cdot u_{0m}) W_0};$$

$$\ldots \quad \ldots \quad \ldots$$

$$I_{y\{\hat{x}(u_m)\}} = \frac{\sum (a_{00} + a_{01} \cdot u_{01} + \ldots + a_{0m} \cdot u_{1m}) W_0}{\sum (a_{00} + a_{01} \cdot u_{01} + \ldots + a_{0m} \cdot u_{0m}) W_0};$$

б) систему индексов, измеряющих влияние на y изменений интенсивности связей между x и u (a_1):

$$I_{y\{\hat{x}(a_{01})\}} = \frac{\sum (a_{00} + a_{11} \cdot u_{11} + \ldots + a_{0m} \cdot u_{1m}) W_0}{\sum (a_{00} + a_{01} \cdot u_{11} + \ldots + a_{0m} \cdot u_{1m}) W_0};$$

$$\ldots \quad \ldots \quad \ldots$$

$$I_{y\{\hat{x}(a_{01})\}} = \frac{\sum (a_{01} + a_{01} \cdot u_{11} + \ldots + a_{1m} \cdot u_{1m}) W_0}{\sum (a_{00} + a_{01} \cdot u_{11} + \ldots + a_{0m} \cdot u_{1m}) W_0};$$

в) индекс, учитывающий изменение свободного члена уравнения регрессии (a_0):

$$I_{y\{\hat{x}(a_0)\}} = \frac{\sum (a_{10} + a_{11} \cdot u_{11} + \ldots + a_{1m} \cdot u_{1m}) W_0}{\sum (a_{00} + a_{11} \cdot u_{11} + \ldots + a_{1m} \cdot u_{1m}) W_0}.$$

Тогда:

$$I_{y\{\hat{x}\}} = I_{y\{\hat{x}(a_0)\}} \cdot \prod_{(l)} I_{y\{\hat{x}(u_l)\}} \cdot \prod_{(l)} I_{y\{\hat{x}(a_l)\}}.$$

Очевидно, что на основе приведенных формул могут быть получены и соответствующие абсолютные эффекты:

$$\Delta y\{\hat{x}(a_0)\}, \ \Delta y\{\hat{x}(u_1)\}, \ \Delta y\{\hat{x}(a_1)\}.$$

Точно так же может быть проанализировано влияние на факторы, детерминирующие W.

Пример. Покажем применение описанной методики анализа. Предположим, что изучается работа каменноугольных шахт одного треста. В качестве результативного признака выступает среднесменный объем добычи угля (W), который может быть представлен как произведение двух факторов: численности рабочих на подземных работах (N) и среднесменной добычи угля на одного подземного рабочего (V) — $W = N \cdot V$. По данным табл. 13.10 определим, как изменился среднесменный объем добычи угля в целом по тресту в отчетном периоде по сравнению с базисным и как на это изменение повлияло изменение численности подземных рабочих и среднесменной добычи на одного рабочего.

Среднесменный объем добычи угля в целом по тресту увеличился:

Таблица 13.10

Данные каменноугольных шахт одного треста

Номер шахты	Базисный период			Отчетный период			Условный среднесменный объем добычи, т
	среднесменная добыча на одного подземного рабочего, т	число подземных рабочих, чел.	среднесменный объем добычи, т	среднесменная добыча на одного подземного рабочего, т	число подземных рабочих, чел.	среднесменный объем добычи, т	
А	V_0	N_0	W_0	V_1	N_1	W_1	N_1V_0
1	5	80	400	5	90	450	450
2	10	100	1000	12	110	1320	1100
3	9	120	1080	12	120	1440	1080
4	7	110	770	9	130	1170	910
5	6	85	510	7	90	630	540
6	9	80	720	10	85	850	765
7	6	90	540	7	90	630	540
8	11	150	1650	12	135	1620	1485
9	5	80	400	5	85	425	425
10	7	95	665	6	100	600	700
Σ	x	990	7735	x	1035	9135	7995

$$I_{w(1/0)} = \frac{\sum W_1}{\sum W_0} = \frac{9135}{7735} = 1{,}18,$$

объем добычи вырос на 18%, что составило в абсолютном выражении 1400 т.

Численность подземных рабочих в данный период увеличилась, и за счет этого среднесменный объем добычи вырос следующим образом:

$$I_{w(N1/0)} = \frac{\sum N_1 \cdot \sum V_0}{\sum N_0 \cdot \sum V_0} = \frac{7935}{7735} = 1{,}033.$$

Как видим, этот фактор в меньшей степени способствовал росту общего объема добычи. За счет него объем добычи вырос только на 3,3%, или на 260 т.

Основная роль в общем изменении результата принадлежит интенсивному фактору — росту среднесменной добычи угля на одного подземного рабочего:

$$I_{w(V1/0)} = \frac{\sum N_1 V_1}{\sum N_1 V_0} = \frac{9135}{7995} = 1{,}142,$$

т.е. за счет роста среднесменной добычи угля на одного подземного рабочего общий объем добычи вырос на 14,2%, или на 1140 т.

Вычисленные индексы образуют систему индексов:

$$I_{w(N)} \cdot I_{w(v)} = I_w;$$
$$1{,}033 \cdot 1{,}142 = 1{,}18.$$

Как отмечает Г. И. Бакланов, влияние каждого фактора на относительное изменение общей абсолютной величины можно получить, выразив соответствующую абсолютную разность в процентах к общей абсолютной величине в базисном периоде[1]. Вычисляя относительное влияние факторов таким образом, мы получим аддитивное разложение относительного изменения результативного признака. При этом относительная оценка влияния первичного признака (N) будет той же самой (+3,3%), а относительная оценка влияния вторичного признака (V) изменится и составит:

[1] *Бакланов Г. И.* Некоторые вопросы индексного метода. — М.: Статистика, 1972. — С. 15—16.

$$\frac{\sum N_1 V_1 - \sum N_1 V_0}{\sum N_0 V_0} = \frac{1140}{7735} \cdot 100 = 14{,}7\%.$$

Тогда: 3,3% + 14,7% = 18,0%.

Ввиду того, что основная роль в общей динамике объема добычи принадлежит производительности труда — среднесменной добыче одного подземного рабочего, на следующем этапе анализа рассмотрим, за счет каких факторов сложится тот или иной уровень производительности труда и как изменение этих факторов сказалось на величине общего объема среднесменной добычи угля по тресту.

Среднесменная добыча подземного рабочего определяется многими факторами, среди которых можно назвать как характеристики рабочих (стаж, квалификация и т.д.), так и характеристики условий труда (используемая техника, степень механизации производственных процессов и др.) и разрабатываемого угольного пласта (длина лавы, мощность пласта и т.д.). Все эти факторы имеют стохастическую связь со среднесменной добычей рабочего, через нее оказывая влияние на общий объем добычи. Предположим, что из всего множества факторов главными и в отчетном, и в базисном периодах оказались только два: мощность пласта и уровень механизации навалки угля. Данные по этим факторам по каждой из десяти шахт приведены в табл. 13.11.

Таблица 13.11

Факторы среднесменной добычи угля

Номер шахты	Базисный период			Отчетный период		
	среднесменная добыча на одного подземного рабочего, т	мощность пласта, м	уровень механизации навалки угля, %	среднесменная добыча на одного подземного рабочего, т	мощность пласта, м	уровень механизации навалки угля, %
A	V_0	M_0	K_0	V_1	M_1	K_1
1	5	75	48	5	74	45
2	10	108	82	12	108	90
3	9	116	85	12	114	90
4	7	98	82	9	95	95
5	6	91	45	7	88	55
6	9	125	100	10	125	100
7	6	85	35	7	85	45
8	11	14	82	12	115	85
9	5	75	67	5	70	70
10	7	101	64	6	90	65

Таблица 13.12

Результаты расчетов

Период	Средние величины			Средние квадратические отклонения			Коэффициенты вариации, %		
	$\overline{V}$	$\overline{M}$	$\overline{K}$	S_v	S_M	S_k	V_v	V_M	V_k
Базисный	7,5	99,0	69,0	2,01	15,80	19,81	26,8	16,0	28,7
Отчетный	8,5	96,4	74,0	2,73	17,46	19,73	32,1	18,1	26,7

Прежде всего определим средние значения признаков, средние квадратические отклонения и коэффициенты вариации (табл. 13.12).

Сравнение отчетных данных с базисными свидетельствует: о возрастании средних значений тех признаков, которые отражают функционирование шахт (эндогенных); о некотором снижении среднего значения мощности пласта (экзогенного признака). Возросла вариация шахт по величине среднесменной выработки одного подземного рабочего и по мощности пласта, тогда как по уровню механизации навалки угля наметилось некоторое выравнивание данных.

Вычисленные значения коэффициентов парной корреляции указывают на тесную связь между признаками (табл. 13.13).

И в том, и в другом периоде среднесменная добыча рабочего теснее коррелирует с мощностью пласта, нежели с уровнем механизации навалки угля. Однако наметилось некоторое снижение величины r_{vm} при повышении r_{vM}. Сравнение коэффициентов парной корреляции зависимой переменной (V)

Таблица 13.13

Коэффициенты парной корреляции между факторами

Период	r_{vM}	r_{vM}	r_{MK}
Базисный	0,903	0,726	0,819
Отчетный	0,893	0,761	0,780

с независимыми переменными и корреляции последних между собой свидетельствует о колинеарности факторов — их тесной линейной связи. При таком соотношении нецелесообразно построение множественной регрессии, куда бы входили оба названных фактора — и мощность пласта и коэффициент механизированной навалки угля. Поэтому построим парную регрессию, описывающую зависимость среднесменной добычи одного рабочего только от мощности пласта: $\hat{v} = a + b$.

Для базисного периода уравнение парной регрессии:

$$\hat{V}_0 = a_0 + b_0 M_0 = -3{,}885 + 0{,}115 V_0;$$

для отчетного периода:

$$\hat{V}_1 = a_1 + b_1 M_1 = -5{,}0 + 0{,}140 M_1.$$

Так как вариация зависимой переменной превосходит вариацию независимой переменной ($v_v > v_M$), свободный член уравнения регрессии в обоих периодах — отрицательная величина ($a < 0$). Сравнение коэффициентов регрессии b_0 и b_1 показывает, что сила влияния данного фактора на среднесменную добычу рабочего растет, а теснота связи падает ($r_{v_1 M_1} < r_{v_0 M_0}$). Если коэффициент детерминации в базисном периоде составил: $r^2_{v_0 M_0} = 81{,}54\%$, то в отчетном: $r^2_{v_1 M_1} = 79{,}74\%$.

Мощность пласта не входит в жесткодетерминированное выражение общего среднесменного объема добычи угля, которое мы анализировали с помощью индексов. Однако этот фактор также может быть учтен в анализе через регрессию $v = f(M)$ и включение в индексы расчетных значений $\hat{v}$. В этом случае индекс, характеризующий влияние изменения среднесменной добычи одного подземного рабочего на величину общей среднесменной добычи угля, должен быть представлен как произведение трех индексов:

$$I_{w(v)} = \frac{\sum N_1 V_1}{\sum N_1 \hat{V}_1} \cdot \frac{\sum N_1 \hat{V}_1}{\sum N_1 \hat{V}_0} \cdot \frac{\sum N_1 V_0}{\sum N_1 V_0}.$$

Расчетные значения среднесменной добычи угля на одного подземного рабочего и соответствующие расчетные значения общей среднесменной добычи представлены в табл. 13.14.

Подставляя в записанную систему индексов расчетные значения среднесменного объема добычи, получаем:

Таблица 13.14

Расчетные значения среднесменной добычи угля, тонн

Номер шахты	Среднесменная добыча одного рабочего*				Среднесменный объем добычи			
А	$\hat{V}_0$	$\hat{V}_1$	$\hat{V}'$	$\hat{V}''$	$\hat{V}_0 N_1$	$\hat{V}_1 N_1$	$\hat{V}'_{N_1}$	$\hat{V}''_{N_1}$
1	4,9	5,4	4,6	6,5	441	486	414	585
2	8,5	10,1	8,5	11,2	935	111	935	1232
3	9,4	11,0	9,2	12,1	1128	1320	1104	1452
4	7,4	8,3	7,0	9,4	962	1079	910	1222
5	6,6	7,3	6,2	8,4	594	657	658	756
6	10,5	12,5	10,5	13,6	893	1062	892	1156
7	5,9	6,9	5,9	8,0	531	621	531	720
8	9,2	11,1	9,3	12,2	1242	1499	1256	1647
9	4,9	4,8	4,2	5,9	416	408	357	502
10	7,7	7,6	6,5	8,7	770	760	650	870
Итого	x	x	x	x	7912	9003	7607	10142

* $\hat{V}_0 = a_0 + b_0 M_0$; $\hat{V}_1 = a_1 + b_1 M_1$.
$\hat{V}' = a_0 + b_0 M_1$; $\hat{V}'' = a_0 + b_1 M_1$.

$$I_{w(v)} = \frac{9135}{9003} \cdot \frac{9003}{7912} \cdot \frac{7912}{7995} = 1{,}0146 \cdot 1{,}1378 \cdot 0{,}9896 = 1{,}1425.$$

Сопоставление первого и последнего индексов показывает, что базисная регрессия $v = \hat{f}(M)$ точнее описывает исходные данные. Этого следовало ожидать, так как $r_{v_0 M_0} > r_{v_1 M_1}$. Средний из трех индексов отражает динамику среднесменного объема добычи под влиянием мощности пласта. В соответствии с изложенной выше методикой этот индекс можно разложить на частные индексы, отражающие влияние изменения величины мощности пласта:

$$I_{w(v(m))} = \frac{\sum (a_0 + b_0 m_1) N_1}{\sum (a_0 + b_0 m_0) N_1} = \frac{\sum V' N_1}{\sum V_0 N_1}.$$

Изменение силы воздействия этого признака на выработку и соответственно на общий объем добычи:

$$I_{w\{v(b)\}} = \frac{\sum(a_0 + b_1 m_1)N_1}{\sum(a_0 + b_0 m_1)N_1} = \frac{\sum V'' N_1}{\sum V' N_1}.$$

Для увязки этих частных индексов следует ввести корректирующий индекс, отражающий изменение свободного члена уравнения регрессии v по M:

$$I_{w\{v(a)\}} = \frac{\sum(a_1 + b_1 M_1)N_1}{\sum(a_0 + b_1 M_1)N_1} = \frac{\sum V_1 N_1}{\sum V'' N_1}.$$

Все величины, требуемые для расчетов этих индексов, представлены в табл. 13.14. С учетом этого

$$I_{w\{v(M)\}} = \frac{7607}{7912} = 0{,}9614\ (-3{,}86\%),$$

т.е. за счет наблюдаемого в отчетном периоде снижения мощности пласта среднесменная добыча угля сократилась в целом по тресту на 3,86%, или на 305 т. Сокращение мощности пласта происходило, как уже было выявлено, наряду с усилением влияния этого фактора — коэффициент регрессии в отчетном периоде выше, чем в базисном ($b_1 = 0{,}140$, $b_0 = 0{,}115$). Увеличение силы влияния мощности пласта на среднесменную выработку, а через нее на объем добычи характеризует следующий индекс:

$$I_{w\{v(b)\}} = \frac{10142}{7607} = 1{,}333,$$

т.е. за счет роста силы связи общий объем среднесменной добычи вырос на 33,3%, или на 2535 т. Влияние изменения свободного члена уравнения регрессии — параметра a — оценивается следующим индексом:

$$I_{w\{v(a)\}} = \frac{9003}{10142} = 0{,}8876.$$

Этот результат никак не комментируется, как и сам параметр a, он не может быть содержательно интерпретирован.

Рассмотренный пример показывает, что подобный анализ основан на определенной условности. Так, оценку влияния изменения коэффициента регрессии мы проводим при базисном значении свободного члена уравнения, тогда как параметры уравнения регрессии связаны друг с другом. Все они получаются в результате решения одной и той же системы уравнений. То же можно сказать в отношении раздельной оценки изменения значения фактора и силы его влияния. Тем не менее соединение индексного и регрессионного методов обогащает анализ, позволяет ввести в него нефункционально связанные факторы.

Рассмотренная методика анализа позволяет измерить влияние факторов, непосредственно не входящих в жестко-детерминированное выражение результативного признака, не только в целом по совокупности, но и по каждому единичному явлению.

Проведение анализа по отдельным единицам с использованием уравнения регрессии обычно основывается на разложении величины отклонения от общей средней ($y_i - \bar{y}$) на две составляющие ($\hat{y}_i - \bar{y}$) и ($y_i - \hat{y}_i$). Если в уравнение регрессии входят все важные и существенные факторы, от которых зависит величина результативного признака, и коэффициент детерминации близок к единице, то остальные, не включенные в уравнение факторы характеризуют индивидуальные, несущественные особенности, зачастую не имеющие количественного выражения. В этом случае разница ($y_i - \hat{y}_i$) образуется за счет несовпадения интенсивности воздействия на y всех учтенных факторов в условиях данной i-й единицы и средней интенсивности их воздействия, выраженной в величинах коэффициентов регрессии, входящих в расчетное значение $\hat{y}_i$. Это дает право интерпретировать разницу ($y_i - \hat{y}_i$) или отношение $y_i/\hat{y}_i$ как показатель того, как эффективность использования учтенных факторов у i-й единицы соотносится со средней эффективностью их использования. Разница ($\hat{y}_i - \bar{y}$) возникает за счет различия в значениях учтенных факторов для данной i-й единицы и в среднем по совокупности. Такое разложение дает возможность выявить резервы, имеющиеся у каждой отдельной единицы, в частности эффективности использования факторов и в части их уровня.

При анализе взаимосвязей в сочетании с изучением динамики явлений нас интересует в первую очередь не соотношение индивидуального и среднего по совокупности значений результативного признака, а изменение его состояния в отчетном периоде по сравнению с базисным ($y_1 - y_0$). В случае

использования регрессионного анализа эта разница может быть представлена следующим образом:

$$(y_1 - y_0) = [(y_1 - \hat{y}_1) - (y_0 - \hat{y}_0)] + (\hat{y}_1 - \hat{y}_0).$$

Первый член разложения характеризует изменение в величине y, вызванное как изменением влияния тех не учтенных в регрессии факторов, которые не коррелируют с учтенными, так и изменением соотношения индивидуальной и средней силы влияния на y учтенных в регрессии факторов. Второй член этого разложения характеризует изменение в величине y, вызванное изменением в значениях факторов, учтенных в регрессии, и изменением средней силы и воздействия на y.

Продолжая рассматривать наш пример, проведем анализ изменения среднесменной добычи угля, приходящейся на одного подземного рабочего (V), по данным отдельных шахт. Все необходимые величины приведены в табл. 13.15.

Учитывая сравнительно низкие значения отчетного и базисного коэффициентов детерминации ($r_0^2 = 0{,}8154$, $r_1^2 = 0{,}7974$), разница фактической и расчетной величин ($V_i - \hat{V}_i$) выражает не только различия в эффективности использования учтенного фактора — мощности пласта — на данной конкретной шахте по сравнению со средней эффективностью по тресту, но и влияние не учтенных в уравнении регрессии факторов.

Таблица 13.15

Отклонения среднесменной добычи угля

Номер шахты	ΔV	$V_0 - \hat{V}_0$	$V_1 - \hat{V}_1$	$\Delta(V - \hat{V})$	$\hat{V}_1 - \hat{V}_0$	$\Delta\hat{V}(M)$	$\Delta\hat{V}(a, b)$
1	0	0,1	−0,4	−0,5	0,5	−0,3	0,8
2	2	1,5	1,9	0,4	1,6	0	1,6
3	3	−0,4	1,0	1,4	1,6	−0,2	1,8
4	2	−0,4	0,7	1,1	0,9	−0,4	1,5
5	1	−0,5	−0,3	0,9	0,7	−0,4	1,1
6	1	−1,5	−2,5	−1,0	2,0	0	2,0
7	1	0,1	0,1	0	1,0	0	1,0
8	1	1,8	0,9	−0,9	1,9	0,1	1,8
9	0	0,1	0,2	0,1	−0,1	−0,7	0,6
10	−1	−0,7	−1,6	−0,9	−0,1	−1,2	1,1

В среднем среднесменная добыча одного подземного рабочего увеличилась в отчетном периоде по сравнению с базисным на 1 т, мощность пласта снизилась в среднем на 2,6 м. Если бы действовал только этот фактор, то средняя добыча снизилась бы на 0,3 т. Таким образом, весь прирост среднесменной добычи вызван действием прочих факторов.

Увеличение абсолютной величины — свободного члена уравнения регрессии параметра a является следствием снижения тесноты прямолинейной связи между мощностью пласта и среднесменной добычей угля на одного подземного рабочего. Данные табл. 13.15 позволяют определить значимость изменения мощности пласта и прочих факторов в общем изменении величины среднесменной добычи, приходящейся на одного подземного рабочего на каждой шахте. Так, нулевое приращение среднесменной выработки на первой шахте явилось результатом равнодействия отрицательного влияния снижения мощности пласта и других факторов в общей тенденции повышения «съема» угля с 1 м мощности пласта. На второй шахте прирост среднесменной добычи одного рабочего обусловлен, с одной стороны, более эффективным использованием мощности пласта, чем в среднем по тресту, с другой стороны, положительным влиянием изменения прочих факторов, как коррелирующих с мощностью пласта, так и не связанных с ним. Подобные заключения можно сделать по третьей, четвертой и другим шахтам.

Таким образом, введение в жесткодетерминированное уравнение связи величин, найденных на основе уравнения регрессии, позволяет учесть в комплексе как жесткодетерминированные, так и стохастические связи.

В экономическом анализе часто решаются задачи, связанные с изучением средних величин, их уровня и динамики, — какова средняя цена 1 кг ржаного хлеба, говядины, средняя заработная плата в промышленности, в экономике в целом и т.д. Изменение средней величины отражает индекс:

$$I_{\bar{y}_{1/0}} = \bar{y}_1 : \bar{y}_0.$$

По данным отчетного и базисного периодов можно построить регрессии — обязательно с одним и тем же набором объясняющих переменных:

$$\hat{y}_1 = a_1 + b_{11}x_1 + \ldots + b_{1k}x_k;$$

$$\hat{y}_0 = a_0 + b_{01}x_1 + \ldots + b_{0k}x_k.$$

Поскольку, как известно из гл. 9, $\bar{y} = a + b_1\bar{x}_1 + \ldots + b_k\bar{x}_k$, то отчетная и базисная регрессии могут быть использованы для анализа изменения среднего уровня $\bar{y}$.

$$I_{\bar{y}} = \frac{\bar{y}_1}{\bar{y}_0} = \frac{a_1 + b_{11}\bar{x}_{11} + b_{12}\bar{x}_{12} + \ldots + b_{1k}\bar{x}_k}{a_0 + b_{01}\bar{x}_{01} + b_{02}\bar{x}_{02} + \ldots + b_{0k}\bar{x}_k}. \qquad (13.37)$$

Средняя величина $\bar{y}$ может изменяться, во-первых, за счет изменений средних значений переменных $\bar{x}_j$ в отчетном периоде по сравнению с базисным, во-вторых, за счет изменения значений коэффициентов регрессии b_j, в-третьих, за счет изменения значения свободного члена уравнения регрессии a_0.

Соответственно получаем систему индексов:

$$I_{\bar{y}} = \frac{\bar{y}_1}{\bar{y}_0} = \prod_{(j)} I_{\bar{y}(\bar{x}j)} \cdot \prod_{(j)} I_{\bar{y}(bj)} \cdot I_{\bar{y}(a)}. \qquad (13.38)$$

Для того чтобы обеспечить это равенство, нужно принять какое-то правило индексации. Например, в соответствии с уже высказанным положением сначала индексируются все $\bar{x}_j$ при постоянных (базисных) значениях коэффициентов регрессии и свободного члена, затем индексируются коэффициенты регрессии при постоянных (отчетных) средних значениях $\bar{x}_j$, затем индексируется свободный член уравнения регрессии при постоянных (отчетных) значениях, как x_j, так и b_j.

Применим рассмотренную методику к анализу среднесменной добычи угля одним рабочим. Среднесменная добыча одного подземного рабочего: базисная — 7,5 т/чел.; отчетная — 8,5 т/чел. (табл. 13.12). Были построены базисная и отчетная регрессии, описывающие связь между среднесменной добычей (y) и мощностью пласта (x). Подставляя средние значения x и y, получим:

$$-3{,}885 + 0{,}115 \cdot 99{,}0 = 7{,}5 \text{ (т/чел.)};$$

$$-5{,}0 + 0{,}140 \cdot 96{,}4 = 8{,}5 \text{ (т/чел.)}.$$

Измерим, как изменилась среднесменная добыча рабочего и как на нее повлияло изменение средней мощности пласта $\bar{x}$, силы влияния этого фактора на добычу b и корректирующего параметра, т.е. свободного члена уравнения регрессии a:

$$I_{\bar{y}} = 8{,}5 : 7{,}5 = 1{,}133 \ (+13{,}3\%);$$

$$I_{\bar{y}(\bar{x})} = \frac{-3{,}885 + 0{,}115 \cdot 96{,}4}{-3{,}885 + 0{,}115 \cdot 99{,}0} = \frac{7{,}201}{7{,}500} = 0{,}96 \ (-4\%);$$

$$I_{\bar{y}(b)} = \frac{-3{,}885 + 0{,}140 \cdot 96{,}4}{-3{,}885 + 0{,}115 \cdot 96{,}4} = \frac{9{,}611}{7{,}201} = 1{,}335 \ (+33{,}5\%);$$

$$I_{\bar{y}(a)} = \frac{-5{,}0 + 0{,}140 \cdot 96{,}4}{-3{,}885 + 0{,}140 \cdot 96{,}4} = \frac{8{,}5}{9{,}611} = 0{,}884 \ (-11{,}6\%).$$

Таким образом, рост среднесменной добычи угля был обусловлен усилением влияния такого фактора, как мощность пласта, ростом его воздействия на добычу угля на 33,5%. Сама мощность пласта несколько уменьшилась, что привело к снижению среднесменной добычи на 4%. Изменение свободного члена тоже оказало негативное влияние на среднесменную выработку (−11,6%).

Все полученные индексы образуют систему индексов: их произведение равно индексу среднесменной добычи.

13.9. Примеры использования индексов в экономико-статистических расчетах

Практически в любом аналитическом обзоре, публикациях итогов развития экономики страны, региона за месяц, квартал, год, в перспективных расчетах обязательно приводятся индексы. Широкое использование индексов в экономико-статистической практике объясняется свойствами этих показателей: во-первых, взаимосвязью частных и общих индексов, что обеспечивает возможность последовательного агрегирования расчетов — по товарам и товарным группам, по территориям, по стране в целом и т.д.; во-вторых, взаимосвязями между индексами разных показателей — урожайности и

валового сбора, производительности труда и фондовооруженности и т.д.

Зная изменение одного из взаимосвязанных показателей, всегда можно определить расчетным путем изменение другого показателя. Например, по данным отчетности промышленных предприятий района известно, что численность занятых в промышленности сократилась в IV квартале по сравнению с I кварталом на 1,5%, объем промышленной продукции снизился на 3%, средняя заработная плата возросла на 15%. Как изменились производительность труда и фонд зарплаты?

$$I_{\text{п.т}} = 0{,}97 : 0{,}985 = 0{,}985;$$

$$I_{\text{ф.з}} = 1{,}15 \cdot 0{,}985 = 1{,}133.$$

Расчеты показывают, что производительность труда снизилась на 1,5%, хотя заработная плата росла, что привело к увеличению фонда заработной платы на 13,3%.

С помощью индексов измеряют динамику производительности труда. Производительность труда может измеряться либо количеством продукции, вырабатываемой в единицу времени q, либо затратами рабочего времени на единицу продукции t. Причем эти показатели находятся в соотношении $q = 1/t$. Первый из них называют прямым показателем производительности труда, а второй — обратным. Сводный индекс производительности труда определяется как средний из индивидуальных индексов: либо $i_q = q_1/q_0$, либо $i_t = t_0/t_1$ (то, что базисное значение показано в числителе, объясняется обратным характером показателя трудоемкости). Отсюда:

$$I_{\text{произв. труда}} = \frac{\sum i_q(q_1 t_1)}{\sum q_1 t_1} = \frac{\sum i T_1}{\sum T_1},$$

где i — индивидуальные индексы часовой, дневной или месячной производительности труда (по видам продукции);

T_1 — общие затраты времени в отчетном периоде соответственно в человеко-часах, человеко-днях или человеко-месяцах.

В последнем случае в качестве T_1 используется численность рабочих.

Важное значение для анализа и прогноза экономических процессов в стране, для международных сравнений имеет *индекс физического объема промышленной продукции*. Методика

его построения основана на последовательном обобщении данных: индексы для более крупных совокупностей представляют собой средние из составных элементов этих совокупностей. Этим определяется порядок расчета индекса физического объема, который включает:

- определение структурных показателей промышленности по отраслям, которые затем используются в качестве веса при агрегировании индивидуальных индексов в общепромышленный;
- отбор товаров-представителей для каждой отрасли, по которым определяется динамика промышленной продукции в каждой отрасли;
- агрегирование отраслевых индексов в общепромышленный.

В соответствии с международной практикой структура промышленного производства определяется по показателю добавленной стоимости (см., например, табл. 13.16). Доли отраслей в добавленной стоимости всей промышленности используются в качестве весов для отраслевых индексов. Расчет проводится по крупным и средним предприятиям.

«Стандартный» набор товаров-представителей включает профильные для каждой отрасли изделия, занимающие значительный удельный вес в общем объеме промышленного производства. По машиностроению и ряду других отраслей товары-представители отражают выпуск этими отраслями товаров народного потребления. Набор товаров учитывает и качественную дифференциацию продукции, направления ее использования (уголь подразделяется на энергетический и коксующийся, прокат — на сортовой и листовой и т.д.). Всего для построения индекса физического объема промышленного производства используются данные примерно по 400 товарам-представителям в разрезе 120 отраслей и производств. В отраслевых индексах выпуск в натуральном выражении продукции по товарам-представителям обобщается по средним оптовым ценам базисного года:

$$I_{q_1} = \frac{\sum q_1 \bar{p}_0}{\sum q_0 \bar{p}_0}.$$

Сводный индекс промышленного производства равен:

$$I_q = \sum I_{q_i} \cdot w_i,$$

где w_i —удельный вес i-й отрасли по показателю.

Не менее важное значение для социально-экономического анализа и международных сравнений имеет *индекс потребительских цен* (ИПЦ). С 1992 г. на всей территории России осуществляется наблюдение за изменением цен и тарифов,

Таблица 13.16

Структура промышленности Санкт-Петербурга в 1999 г.

Отрасль	Доля отрасли во всей промышленности (по добавленной стоимости по чистым отраслям), %
Промышленность	100
Электроэнергетика	5
Топливная промышленность	0
Черная металлургия	2
Цветная металлургия	1
Химическая и нефтехимическая промышленность (без химико-фармацевтической промышленности)	2
Машиностроение и металлообработка (без промышленности медицинской техники)	49
Лесная, деревообрабатывающая и целлюлозно-бумажная промышленность	3
Промышленность строительных материалов	2
Стекольная и фарфоро-фаянсовая промышленность (без предприятий по производству медицинских изделий из стекла и фарфора)	0
Легкая промышленность	2
Пищевая промышленность	27
Микробиологическая промышленность	0
Мукомольно-крупяная и комбикормовая промышленность	2
Медицинская промышленность	2
Полиграфическая промышленность	2
Другие промышленные производства	1
Государственная приемка продукции в промышленности, государственный надзор и контроль за стандартами и средствами измерений	0
Хозяйственное управление промышленностью	0

которое ведет специально созданная Госкомстатом государственная служба.

Вторым источником информации служат данные бюджетной статистики. Примерно 45 тыс. домохозяйств в России ведут подробный учет своих доходов и расходов.

На основе этих двух информационных потоков проводится расчет ИПЦ по фиксированному набору основных потребительских товаров и услуг по методологии, принятой в международной практике.

Индекс потребительских цен измеряет изменение стоимости фиксированной потребительской корзины товаров и услуг, используемых семьями. Корзина товаров и услуг фиксирована с тем, чтобы данному уровню жизни соответствовало одно и то же значение индекса. При таком подходе изменения ИПЦ могут вызываться только изменением цен, но не переменами в структуре потребления в результате изменения доходов или появления новых товаров. По этой причине ИПЦ называют *индексом стоимости жизни*. Он широко используется в качестве показателя инфляции.

Национальный ИПЦ рассчитывается на основе данных по 266 крупным городам России, представляющим все федеральные округа. Каждый из этих городов имеет население более 200 тыс. человек, в их число входят 13 городов-миллионеров. В сумме население отобранных городов составляет примерно $1/3$ городского населения Российской Федерации. Информация о ценах, собранная по этим городам, применяется для расчета средних цен с использованием в качестве весов суммы расходов всех домохозяйств каждого города. На основе этих данных строятся и региональные ИПЦ и, если необходимо, для отдельных товаров и товарных групп.

Общегосударственный ИПЦ рассчитывается на основе отношений цен на 410 товаров и услуг, зарегистрированных в 266 городах. Для каждого города отношения цен агрегируются в общегосударственные средние с использованием общих расходов в каждом городе в качестве весов (численность населения города умножается на среднедушевое потребление, данные о котором берутся из бюджетного обследования).

Расчетная формула ИПЦ:

$$I_{t/0} = \frac{\sum\limits_{(j)} \left[p_{0j} Q_{0j} \frac{p_{tj}}{p_{0j}} \right]}{\sum\limits_{(j)} p_{0j} Q_{0j}} \cdot 100\%, \qquad (13.39)$$

где p_{0j} — цена товара j в базисном периоде;
Q_{0j} — количество товара j в базисном периоде;
p_{tj} — цена товара j в периоде t.

Очевидно, что эта формула тождественна формуле индекса цен Ласпейреса. Ее можно представить как

$$I_{t/0} = \sum_{(j)} w_{0j} i_{0j} \cdot 100, \qquad (13.40)$$

где $w_{0j} = \frac{p_{0j} Q_{0j}}{\sum p_{0j} Q_{0j}}$ — доля расходов на товар j в общих расходах.

Однако практически трудно использовать и первое, и второе выражение ИПЦ, так как оба варианта включают отношение цены для периода t к цене в базисный период (p_t/p_0) и предполагают сравнение изменений цен для каждого товара за длительные периоды с сохранением характеристик данных товаров. Эти условия трудно выполнить при изменении круга продаваемых товаров, замещении товаров, изменении структуры товарных потоков.

Поэтому применяется вариант ИПЦ с использованием отношения цены товара в периоде t к цене в предыдущем периоде $t-1$ (p_t/p_{t-1}):

$$I_{t/0} = \frac{\sum\limits_{(j)} (p_{tj}/p_{t-1,j}(p_{t-1} Q_{0j}))}{\sum\limits_{(j)} p_{0j} Q_{0j}} \cdot 100, \qquad (13.41)$$

где

$$p_{t-1,j} Q_{0j} = p_{0j} Q_{0j} \frac{p_{1j} p_{2j}}{p_{0j} p_{1j}} \ldots \frac{p_{t-1j}}{p_{t-2j}}.$$

Последняя формула ИПЦ тождественна двум предыдущим, но использование цепных сравнений цен облегчает введение новых товаров или их замещение, когда возникает такая необходимость.

Индекс потребительских цен строится путем последовательного агрегирования данных. Сначала определяются потоварные индексы цен, охватывающие все виды торговли, затем — индексы цен по товарным группам, после чего строится сводный ИПЦ.

Например, в состав ИПЦ входит индекс потребительских цен на мясо и мясопродукты:

$$I_{p\ \text{мясо и мясопродукты}} = \frac{\sum I_j w_0}{\sum w_0},$$

где w_0 — удельный вес расходов на покупку данных товаров в потребительских расходах населения (по данным бюджетных обследований).

В табл. 13.17 приведены в качестве примера индексы цен на мясо и мясопродукты в Санкт-Петербурге за 2002 г. Доля этой группы товаров в потребительских расходах населения составляла в указанном году 0,15141.

Каждый из индексов цен по товарам этой группы обобщает динамику цен на данный товар по видам торговли.

Общенациональный ИПЦ строится как средний из территориальных индексов, взвешенных по численности населения:

$$I_{\text{ПЦ}} = \frac{\sum I_{\text{ПЦ}} \cdot \text{численность населения}}{\sum \text{численность населения}}.$$

Трудно перечислить все индексы, используемые в социально-экономической статистике. Это и индексы урожайности, структуры посевных площадей, валового сбора, и индексы себестоимости продукции, рентабельности и т.д. В условиях инфляции особенно большое значение приобретают индексы цен. Кроме индекса потребительских цен службы государственной статистики рассчитывают индексы оптовых цен (цен производства) и др. Индексы цен выполняют роль дефлятора, т.е. используются для пересчета показателей, выраженных в текущих ценах, в базисные цены, т.е. в цены года, принятого в качестве базисного. С помощью дефляторов исчисляется динамика сводных статистико-экономических по-

Таблица 13.17

Данные по товарной группе «мясо и мясопродукты» в Санкт-Петербурге (декабрь 2002 г., к декабрю 2001 г., в процентах)

Товар	Индексы цен
Говядина	100,3
Баранина	101,1
Свинина	95,5
Мясо птицы	101,0
Мясные полуфабрикаты	100,3
Пельмени	100,3
Субпродукты	95,2
Итого мясопродуктов	100,7
ИПЦ	114,7

казателей — валового внутреннего продукта, валового национального продукта, объема капитальных вложений и т.д. С помощью ИПЦ решаются вопросы индексации доходов населения. В практических расчетах строятся как изолированные индексы, так и системы взаимосвязанных индексов. На их основе проводится анализ изменения сложных явлений по факторам. Однако, проводя аналитические расчеты с помощью индексов, помните, что строгость их формул, взаимные увязки, количественные оценки (относительные и абсолютные) вкладов отдельных факторов в совокупное изменение нельзя воспринимать как абсолютную истину. Это всего лишь приближение к истине, которое получено при той или иной методике построения индексов (система выбора весов, базы сравнения, построения исходного уравнения связи между признаками). Не обольщайтесь кажущейся точностью, отнеситесь к результатам критически!

Большое значение в экономической практике имеют соотношения в изменениях показателей, т.е. соотношения между величинами их индексов. Например, известно, что в эффективной экономике темпы роста производительности труда должны опережать темпы роста заработной платы:

$$I_{\text{произв. труда}} > I_{\text{заработной платы}}.$$

Инвестиции в знание должны расти быстрее, чем в основные фонды: в США +3,4% против +2,2%.

Для развития предприятий оптимально следующее соотношение динамики основных показателей:

$$I_{\text{балансовой прибыли}} > I_{\text{реализации}} > I_{\text{авансированного капитала}} > 100\%.$$

Такого рода соотношения принято называть «экономической нормалью» или «динамическим нормативом».

Сравнение с нормалью используется и в аудиторской деятельности для заключения о финансовом положении предприятия, его потенциале. Например, для трудоемкого производства в качестве нормали формулируется следующее неравенство:

$$I_{\text{объема реализации}} > I_{\text{материальные затраты на производство}} > I_{\text{численность промышленно-производственного персонала}} > I_{\text{средняя стоимость основных производственных фондов}}.$$

Для того чтобы определить соответствие фактической динамики нормали, нужно иметь данные об изменении показателей (индексы) за несколько периодов. Например, оказалось, что поквартальные индексы за два года показывают следующее (табл. 13.18).

Только в трех кварталах соотношение в изменении показателей было близко к нормали. Аудитор обязан указать на это в своем заключении и рекомендовать менеджерам обратить внимание на причины: высокие цены поставщиков, из-

Таблица 13.18

Значения индексов

Индекс	1-й год				2-й год			
	I	II	III	IV	I	II	III	IV
I объема реализации	x	0,994	0,985	1,041	1,055	0,990	1,002	1,036
I материальных затрат на производство	x	0,955	0,967	1,096	1,007	0,960	0,983	1,056
I численности промышленно-производственного персонала	x	0,972	0,995	1,016	1,000	0,989	1,009	0,998
I стоимости основных производственных фондов	x	1,001	0,999	1,001	1,001	1,000	1,002	1,006

быточная численность персонала, неэффективная структура и использование основных фондов и т.д.

Динамика, соответствующая экономической нормали, обычно определяет стратегию развития предприятий, и для управления компанией (фирмой) важно проводить сравнение фактического соотношения темпов изменения показателей с «нормальным», выявлять, в каком звене нормали возникли нарушения, и вносить коррективы в деятельность предприятия.

РЕЗЮМЕ

Слово «индекс» означает показатель. В статистике индексы используются в качестве показателей изменений. Индекс — это показатель сравнения двух состояний одного и того же явления (простого или сложного, состоящего из соизмеримых *или несоизмеримых элементов*).

Индексы измеряют изменения сложных явлений. С их помощью можно не только дать обобщенную оценку изменения, но и выявить роль отдельных факторов.

Индексы являются показателями сравнения как с прошлым периодом, так и с другой территорией, а также с некоторым нормативом или плановым заданием.

Каждый индекс включает отчетные и базисные данные.

Сравнение с отдаленной базой может быть проведено непосредственно с помощью базисного индекса, охватывающего весь период, или поэтапно — с помощью цепных индексов.

Индексы подразделяются на сводные (общие) и индивидуальные.

Каждый сводный индекс может быть представлен как средний из индивидуальных. В этом смысле, как и любая средняя, сводный индекс характеризует центральную тенденцию. Значение индекса среднего из индивидуальных зависит от изменений осредняемых индивидуальных индексов и от изменений признака-веса.

Агрегатные индексы считаются основной формой индексов. Они выполняют две функции — синтетическую и аналитическую. С точки зрения последней аналитические индексы должны образовывать систему индексов. Это требование налагает определенные ограничения на построение каждого

аналитического индекса, входящего в одну и ту же систему: каждый из них должен использовать веса разных периодов. Так, индекс Ласпейреса строится на весах базисного периода, а индекс Пааше — на весах отчетного периода.

Аналитический индекс включает индексируемый признак, изменение которого характеризует данный индекс, и признак-вес.

Соотношение аналитических индексов с весами разных периодов позволяет измерить эффект совместного изменения изучаемых признаков. Эта составляющая аналитического разложения имеет особое значение, если все взаимосвязанные индексы строятся на весах одного и того же периода.

Индексы считаются правильно построенными, если они удовлетворяют ряду тестов:

- обратимости во времени;
- обратимости по факторам;
- «кружному» испытанию;
- соизмеримости;
- пропорциональности;
- включения — исключения.

Индексы широко используются для анализа изменений средних взвешенных величин (средней заработной платы, производительности труда, трудоемкости и т.д.). С этой целью применяется система индексов: индекс переменного состава, индекс постоянного состава, индекс структуры.

Индекс переменного состава — это сравнение отчетного и базисного значений взвешенной средней.

Индекс постоянного состава измеряет, как изменяется осредняемый признак (при постоянстве признака-веса).

Индекс структуры измеряет, как изменилась величина средней за счет признака-веса (при постоянстве осредняемого признака).

Построение индексов и характер решаемых ими задач зависят от уровня обобщения данных:

- по элементам признака в рамках одной единицы совокупности ($n = 1$, $m > 1$);
- по группе единиц в рамках одного элемента ($n > 1$, $m > 1$);
- по группе единиц и группе элементов ($n > 1$, $m > 1$).

В последнем случае данные следует обобщать по элементам, а затем — по всем единицам.

Использование индексов для решения аналитических задач возможно при условии жесткодетерминированной связи признаков — либо мультипликативной, либо аддитивной.

Переход от одного уровня анализа (жесткодетерминированные связи) на другой (стохастические связи) возможен путем введения уравнений регрессии в индекс и последовательной оценки изменений объясняющих переменных и параметров уравнений регрессии.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Адамов В. Е.* Факторный индексный анализ. Методология и проблемы. — М.: Статистика, 1977.
2. *Аллен Р.* Экономические индексы: Пер. с англ. — М.: Статистика, 1980.
3. *Зоркальцев В. М.* Индексы цен и инфляционные процессы. — Новосибирск: Наука — Сибирская издательская фирма РАН, 1996.
4. *Казинец Л. С.* Теория индексов. — М.: Госстатиздат, 1963.
5. *Ковалевский Г. В.* Индексный метод в экономике. — М.: Финансы и статистика, 1989.
6. *Фишер И.* Построение индексов: Пер. с англ. — М.: Изд-во ЦСУ СССР, 1928.

Глава
14

СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ СТРУКТУРЫ СОВОКУПНОСТИ И ЕЕ ИЗМЕНЕНИЙ

14.1. Показатели простой (одномерной) структуры

Развитие статистической совокупности проявляется не только в количественном росте или уменьшении элементов этой системы, но также и в изменении ее структуры.

Структура — это строение, форма организации системы, состоящей из отдельных элементов и связей между ними. Так, человеческий организм представляет собой систему, состоящую из различных тканей, органов, закономерным образом взаимодействующих друг с другом. Экономика страны состоит из отраслей, предприятий, учреждений, связанных не только материально, но и информационно, энергетически.

Решающим условием дальнейшего развития человеческого общества в современную эпоху стало не простое расширение, количественное возрастание его параметров (численности населения, объемов производства и потребления ресурсов), а структурное изменение — переход от ресурсопотребляющей к ресурсосберегающей стратегии. На смену экспоненциальному росту потребления невозобновляемых ресурсов приходят экологически ориентированное производство, ограничение роста населения как условия повышения качества жизни.

Соответственно возрастает роль методов и показателей статистики, характеризующих структуру социальных, производственных, технологических систем, ее изменений.

В подразд. 3.2 были рассмотрены относительные показатели, характеризующие простую (одномерную) структуру: доля или удельный вес отдельных элементов в итоге абсолютного признака совокупности. В гл. 5 рассмотрены система показателей и методика анализа распределения совокупности по значениям какого-либо отдельного признака. В данном разделе рассмотрены особенности изучения структуры по такому признаку, который способен принимать как положительные, так и отрицательные значения (например, финансовый результат деятельности фирмы, предприятия, сальдо миграции).

Примером такого рода являются данные табл. 14.1. Половину полученной прибыли обеспечило растениеводство; значительна доля в прибыли продукции промышленной переработки сельскохозяйственной продукции. Показатели графы 3 говорят о структуре убытков, указывая на неблагополучное

Таблица 14.1

Структура финансового результата деятельности сельскохозяйственных предприятий района

Вид деятельности	Прибыль (+), убыток (−), млн руб.	В % к итогу			Модули сумм, млн руб.	В % к сумме модулей
		прибыль	убыток	сальдо финансового результата		
А	1	2	3	4	5	6
Растениеводство	+150	50	—	+75	150	37,5
Животноводство	−40	—	40	−20	40	10,0
Промышленная переработка	+120	40	—	+60	120	30,0
Услуги на сторону	+21	7	—	+10,5	21	5,25
Жилищно-коммунальное хозяйство	−60	—	60	−30	60	15,0
Прочая деятельность	+9	3	—	+4,5	9	2,25
Итоги: Прибыль	+300	100	—	—	400	100,0
Убыток	−100	—	100	—		
Сальдо	+200	—		100		

финансовое состояние жилищно-коммунального хозяйства. По данным графы 4 рассчитаны показатели структуры знакопеременного признака — финансового результата. Эти показатели структуры также имеют разные знаки. Никакого запрета на отрицательную величину долей статистика не налагает; требуется, чтобы сумма долей была равна 100% и только. Экономический смысл показателей структуры финансового результата совершенно очевиден: растениеводство дало 50% прибыли, но не 50, а 75% финансового результата от всех видов деятельности. Животноводство снизило финансовый результат не на 40% данного им убытка, а на 20%. Знакопеременные доли в графе 5 реально отражают «вклады» каждой из отраслей деятельности в конечный общий финансовый результат. Другой способ получить аналогичные по экономическому значению показатели — построение показателей структуры по модулям финансовых результатов. В этом случае нужно, отбросив знаки, сложить прибыли и убытки, а затем вычислить доли каждой отрасли. Показатели графы 6 абстрагированы от знака, они характеризуют не направление, а только сравнительную величину влияния, «вклада» каждой отрасли в образование финансового результата. Ранжируя отрасли по этим долям, можно сделать вывод, что самое большое влияние оказало растениеводство, на втором месте — промышленная переработка, на третьем по силе влияния — жилищно-коммунальное хозяйство и т.д. Показатели графы 6 пропорциональны показателям графы 4, но последние, кроме того, характеризуют и направление «вклада» отраслей, а потому более информативны; именно им нужно отдать предпочтение.

14.2. Показатели иерархической (древовидной) структуры

Иерархической (древовидной) структурой называется сложная структура, образуемая при последовательном дроблении системы на все более однородные группы элементов. Она состоит из нескольких уровней («шагов» дробления). Таковы, например, административно-управленческая структура предприятия — двух-трехзвенная или более сложная структура народного хозяйства по крупным отраслям, подотраслям и

группам однородных предприятий; структура товарооборота по группам товаров, их видам, сортам, размерам и т.д.

Рассмотрим пример иерархической структуры (рис. 14.1)[1].

На этом рисунке указаны шесть уровней и площадь каждой иерархии (дробления), доля этой площади в общей величине земельной площади хозяйства. Отметим, что вся иерархическая структура насчитывает шесть уровней, не считая нулевого, на котором еще нет дробления.

Иерархическая структура характеризуется не только долями объема признака, но и дополнительными показателями.

1. Характеристикой степени сложности структуры:

а) числом уровней дробления («порядок» структуры). На рис. 14.1 приведена структура шестого порядка.

2. Средним порядком структуры, т.е. средним номером уровня, взвешенным по долям объема признака, дробление которых завершилось на данном уровне:

$$\overline{\Pi} = \sum_{i=1}^{k} \Pi_i \cdot d_i, \qquad (14.1)$$

где $\overline{\Pi}$ — средний порядок;
k — число уровней;
Π_j — номер уровня (порядок дробления);
d_i — доля признака на i-м уровне.

По данным рис. 14.1 $\overline{\Pi} = 6 \cdot 0{,}08 + 5 \cdot 0{,}23 + 4 \cdot 0{,}13 + 3 \cdot 0{,}4 + 2 \cdot 0{,}16 = 3{,}67$.

Эта величина характеризует среднее число дроблений объема признака.

3. Общим числом конечных (т.е. не дробящихся далее) ветвей структуры. В данном примере имеем 28 конечных ветвей.

4. Средним числом конечных ветвей, приходящихся на один уровень:

$$\overline{b} = \frac{\sum_{i=1}^{k} b_i}{k}. \qquad (14.2)$$

В примере $\overline{b} = \frac{28}{6} = 4{,}67$.

[1] *Агапова Т. Н.* Методы статистического изучения структуры сложных систем и ее изменения. — М.: Финансы и статистика, 1996. — С. 59—62.

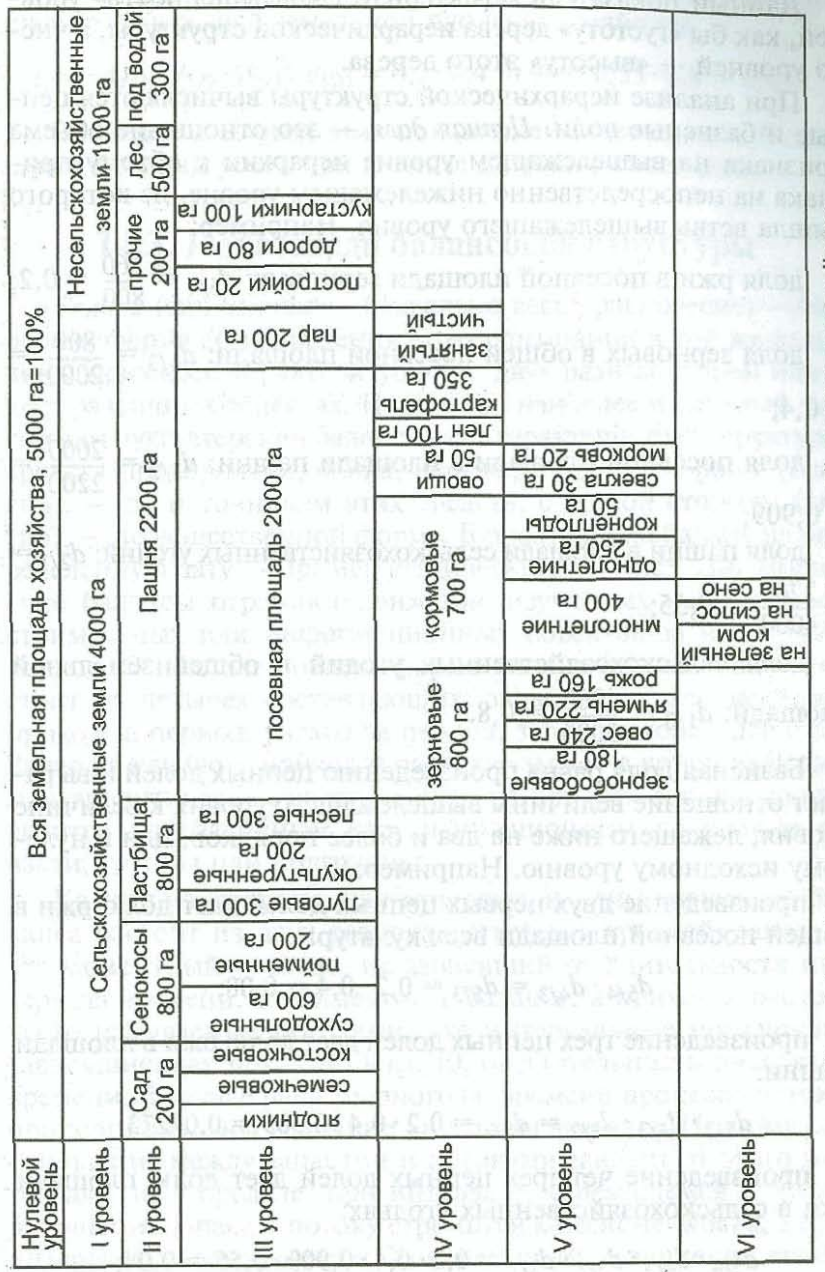


Рис. 14.1. Иерархическая структура земельной площади хозяйства

Данный показатель характеризует «насыщенность» уровней, как бы «густоту» дерева иерархической структуры, а число уровней — «высоту» этого дерева.

При анализе иерархической структуры вычисляются цепные и базисные доли. *Цепная доля* — это отношение объема признака на вышележащем уровне иерархии к объему признака на непосредственно нижележащем уровне, из которого вышла ветвь вышележащего уровня. Например:

доля ржи в посевной площади зерновых: $d_{5/4} = \frac{160}{800} = 0{,}2$;

доля зерновых в общей посевной площади: $d_{4/3} = \frac{800}{2000} = 0{,}4$;

доля посевной площади в площади пашни: $d_{3/2} = \frac{2000}{2200} = 0{,}909$;

доля пашни в площади сельскохозяйственных угодий: $d_{2/1} = \frac{2200}{4000} = 0{,}55$;

доля сельскохозяйственных угодий в общей земельной площади: $d_{1/0} = \frac{4000}{5000} = 0{,}8$.

Базисная доля равна произведению цепных долей и выражает отношение величины вышележащего уровня к величине уровня, лежащего ниже на два и более порядков, или к нулевому исходному уровню. Например:

произведение двух первых цепных долей дает долю ржи в общей посевной площади всех культур:

$$d_{5/4} \cdot d_{4/3} = d_{5/3} = 0{,}2 \cdot 0{,}4 = 0{,}08;$$

произведение трех цепных долей дает долю ржи в площади пашни:

$$d_{5/3} \cdot d_{4/3} \cdot d_{3/2} = d_{5/2} = 0{,}2 \cdot 0{,}4 \cdot 0{,}909 = 0{,}07273;$$

произведение четырех цепных долей дает долю площади ржи в сельскохозяйственных угодьях:

$$d_{5/4} \cdot d_{4/3} \cdot d_{3/2} \cdot d_{2/1} = 0{,}2 \cdot 0{,}4 \cdot 0{,}909 \cdot 0{,}55 = 0{,}04;$$

Наконец, произведение всех цепных долей дает долю площади ржи в общей земельной площади хозяйства:

$$d_{5/4} \cdot d_{4/3} \cdot d_{3/2} \cdot d_{2/1} \cdot d_{1/0} = 0{,}2 \cdot 0{,}4 \cdot 0{,}909 \cdot 0{,}55 \cdot 0{,}8 = 0{,}032.$$

Очевидна аналогия этих показателей с цепными и базисными темпами роста при анализе динамического ряда.

14.3. Показатели балансовой структуры

Баланс (фр. balance — буквально весы, равновесие) — это особая форма сопоставления структуры одной и той же величины признака, характеризуемой с двух разных сторон или в двух различных аспектах. Например, наиболее известный читателям бухгалтерский баланс — это характеристика структуры средств предприятия, банка, фирмы, с одной стороны (пассив), — по источникам этих средств, с другой стороны (актив), — по вещественной форме. Бухгалтерский баланс на определенную дату — пример статического баланса. Динамические балансы отражают движение изучаемых натуральных, стоимостных или информационных объектов за некоторый период. В наиболее общей форме динамический баланс состоит из четырех составляющих: запас на начало периода, приход за период, расход за период, запас на конец периода. Запас на начало + приход = расход + запас на конец периода. Для аналитических целей каждая из четырех составляющих делится по различным классификационным признакам на части, группы или подгруппы.

Каждая из «сторон», или «половинок», динамического баланса состоит из двух разнокачественных уровней: запас — это моментный уровень, не зависящий от длительности интервала времени, отражаемого в балансе, а приход и расход, часто называемые потоками, это интервальные показатели, зависящие, как показано в гл. 10, от длительности интервала времени. В случае равномерного во времени процесса потоки пропорциональны величине интервала времени. Поэтому соотношение между запасами и потоками зависит от этого интервала, и, в пределе, при интервале, стремящемся к нулю, отношение запаса к потоку стремится к бесконечности, а при интервале, стремящемся к бесконечности, отношение запаса

к потоку стремится к нулю. Данное свойство непременно должно учитываться при анализе балансовых структур.

Но при заданной величине интервала времени, например, равной одному году, отношение запаса к потоку является очень существенным структурным показателем, характеризующим изучаемый объект. Если запас значительно превышает величину потока за год, объект можно условно назвать консервативным. Таковы, как правило, основные фонды предприятия. Их поступление за год и выбытие за год обычно не достигают и 50% запаса, т.е. наличия на 1 января, или среднегодового. Напротив, если поток за год существенно превышает запас, объект можно условно назвать мобильным. Таковы оборотные средства предприятий, товары в розничной торговле, денежные средства большей части населения.

Остановимся на показателях соотношения между запасом и потоком. Примем такой вариант, когда показателем запаса считается его среднегодовой уровень (средняя из величин запаса на начало и конец года или точнее — хронологическая средняя из данных на начало каждого месяца или квартала, как показано в гл. 12), а величиной потока будем считать меньшую из величин входящего потока (поступление) и исходящего потока (выбытие). Это допущение позволяет отделить поток, проходящий через объект, от прироста или уменьшения запаса за год.

Пример. Пусть начальный запас данного материала составил 2000 ед., приход за год — 5000 ед., выбытие — 4500 ед., конечный запас — 2500 ед. Тогда среднегодовой запас составит 2250 ед., поток — 4500 ед. Отношение потока к среднегодовому запасу равно двум. Эту величину обычно интерпретируют как число оборотов данного материала за год, т.е. величина имеет единицу измерения «год в минус первой степени», что и вытекает из отношения:

$$\frac{\text{Поток} = 4500 \text{ ед./год}}{\text{Запас} = 2250 \text{ ед.}} = 2 \text{ год}^{-1}, \text{ т.е. два оборота в год.}$$

Если предположить, что поток был в течение года равномерным, то за квартал он составит 1125 ед., тогда средний запас за I квартал составит:

$$\frac{2000 + \left(2000 + \frac{500}{4}\right)}{2} = 2062{,}5 \text{ ед., а число оборотов:}$$

$$\frac{1125 \text{ ед./квартал}}{2062{,}5 \text{ ед.}} = 0{,}545 \text{ оборота в квартал,}$$

или 0,545 оборота квартал$^{-1}$.

За IV квартал имеем средний запас:

$$\frac{\left(2000 + 3\frac{500}{4}\right) + 2500}{2} = 2437{,}5 \text{ ед.,}$$

а число оборотов: 1125 ед./квартал : 2437,5 ед. = 0,462 квартала$^{-1}$. Как видим, при равномерном потоке и росте запаса отношение потока к запасу постепенно уменьшается, скорость оборота замедляется. При равномерном потоке и сокращении запаса, наоборот, скорость оборота будет возрастать. Обратная величина — отношение запаса к потоку за год составит:

$$\frac{2250 \text{ ед.}}{4500 \text{ ед./год}} = 0{,}5 \text{ года, или } 182{,}5 \text{ дня.}$$

При указанных выше единицах измерения прямого и обратного показателей их произведение равно единице. В нашем примере при двух оборотах в год средства можно считать умеренно мобильными.

Конечно, изучение структуры динамического баланса не ограничивается приведенными общими показателями. Значительный интерес представляет изучение структуры входящего и исходящего потоков, например долей импорта и собственного производства в приходе товара, реализации и потерь в исходящем потоке товаров и других отношений. Поскольку они выражаются обычными долями, нет необходимости рассматривать методику их определения.

Одним из важнейших следствий деятельности человечества на планете Земля в настоящее время является возрастание содержания в атмосфере окиси углерода. В результате увеличивается «парниковый эффект» атмосферы, повышается сред-

Таблица 14.2

Годовой баланс CO_2 в атмосфере Земли, млрд т

Наличие на начало года и поступления за год		Выбытие за год и наличие на конец года	
Наличие на начало года	1540	Выбытие	
Поступления:			
Дыхание растений и животных	220	Фотосинтез растений на суше	440
выделяется из почв	220	Поглощение в океане	367
выделяется из океана	341	в том числе осаждение	
от сгорания лесов	5	в известняке	11
выбросы промышленности, транспорта, бытовые	37		
Итого поступлений	823	Итого выбытий	807
		Наличие на конец года	1556
Баланс	2363	Баланс	2363

Источник. Добровольцев Г. В., Куст Г. В. Деградация почв — «Тихий кризис планеты» // Природа. — 1996. — № 10. — С. 53—63.

няя температура воздуха, что может привести к очень серьезным и неблагоприятным для человечества последствиям. Рассмотрим структуру динамического баланса содержания углекислого газа в атмосфере (табл. 14.2).

Отношение «потока» — величины выбытия к среднегодовому запасу составляет 0,521, что характеризует систему как весьма мобильную. Доля антропогенных выбросов в поступлении CO_2 невелика, только 4,5%. Однако быстрый рост антропогенных выбросов в поступлении CO_2 в XIX в. и особенно в XX в. привел к превышению его выбытия. При сохранении выбросов на нынешнем уровне запас содержания CO_2 в атмосфере возрастает примерно на 1% в год, что ведет к удвоению доли CO_2 среди всех компонентов атмосферы за столетие и резкому возрастанию «парникового эффекта», так как именно молекулы CO_2 (а также метана) задерживают уходящее с поверхности Земли низкочастотное тепловое излучение. Таким образом, казалось бы небольшое нарушение структуры баланса за достаточно длительное время может привести к очень крупным изменениям системы.

Перейдем к специфическим показателям, характеризующим структурные соотношения между различными сторонами бухгалтерского баланса.

Таблица 14.3

Баланс ОАО АКБ «Автобанк»

Статья баланса	На 01.01.1998 г.		На 01.01.2000 г.	
	млн. руб.	доля в валюте баланса, %	млн. руб.	доля в валюте баланса, %
АКТИВ				
1. Остатки на счетах в Центральном банке, касса	1372	14,34	970	8,47
2. Средства в кредитных организациях	77	0,80	228	1,99
3. Вложения в ценные бумаги	2984	31,18	1750	15,28
4. Чистые кредиты и лизинг клиентам	4253	44,44	7905	69,03
5. Основные средства и нематериальные активы	325	3,40	331	2,89
6. Прочие активы	559	5,84	268	2,34
Всего активов	9570	100	11452	100
ПАССИВ				
I. Собственные средства, в том числе:				
1. Уставный капитал	444	4,64	972	8,49
2. Акции и другие собственные источники	1212	12,67	232	2,03
3. Нераспределенная прибыль отчетного года	230	2,40	152	1,33
Итого собственных источников	1886	19,71	1356	11,85
II. Обязательства, в том числе:				
4. Кредиты Центрального банка	—		2508	21,90
5. Средства кредитных организаций	2889	30,19	2849	24,87
6. Средства клиентов	3034	31,70	3382	29,53
7. Выпущенные долговые обязательства	1410	14,74	1119	9,77
8. Прочие обязательства	98	1,02	238	2,08
Итого обязательств	7431	77,65	10096	88,15
9. Прочие пассивы	253	2,64	—	—
Всего пассивов	9570	100	11452	100

Пример. Рассмотрим баланс без подробного состава статей акционерного коммерческого банка «Автобанк» (табл. 14.3).

В анализе структуры баланса применяются, кроме долей отдельных статей в итоге валюты баланса, еще и такие относительные показатели, которые измеряют отношения между статьями актива и статьями пассива. К структурным характеристикам (долям), выражающим существенные характеристики банка, относится, например, доля собственных средств в итоге пассива. Так, на 01.01.1998 г. доля собственных источников средств «Автобанка» составила: 1886 : 9570 = = 19,71%, а на 01.01.2000 г. — 1356 : 11952 = 11,85%. Значительное уменьшение доли собственных средств банка на 7,85 пункта, или на 39,9%, означает ухудшение устойчивости банка, произошедшее после финансового кризиса 1998 г.

Примером второго рода показателей структуры баланса может служить отношение суммы средств, которыми банк может располагать, к сумме обязательств, которые могут быть предъявлены клиентами (вкладчиками) к оплате. Эти показатели называют показателями ликвидности. Показатель «быстрая ликвидность» может быть получен (при той степени подробности статей, какая имеется в табл. 14.3) как отношение статьи 1 актива к статьям 5 и 6 пассива. На 01.01.1998 г. этот показатель составил: 1372 : 5923 = 0,232, или 23,2%. Конечно, вряд ли все клиенты одновременно предъявят требования к оплате, поэтому в спокойной обстановке показатель «быстрая ликвидность» можно считать достаточным. Но в случае банковского кризиса, паники среди вкладчиков этот показатель явно недостаточен. На 01.01.2000 г. показатель «быстрая ликвидность» составил: 970 : 6231 = 0,1557, или 15,57%, что значительно хуже предыдущего значения.

Общий показатель покрытия обязательств есть отношение всех активов, кроме неликвидных (основных средств и прочих активов), ко всем обязательствам. На 01.01.2000 г. этот показатель, вычисленный на основе долей, составил: (100% − 2,89% − 2,34%) : 88,15% = 1,075, или 107,5%. Это означает, что в долгосрочном периоде банк в состоянии расплатиться с кредиторами, если его должники вернут банку долги.

Нормальная деятельность банка (предприятия) требует, чтобы величина общего коэффициента покрытия обязательств превышала 1, а лучше, если она имеет значение от 1,5 до 20,0.

14.4. Показатели многомерной структуры с пересекающимися признаками

Если общий объем признака подразделен по одному группировочному признаку, а затем каждый групповой и общий объемы снова подразделены по другому группировочному признаку, то образуется *многомерная*, в простейшем случае — двухмерная *структура с пересекающимися признаками.*

Пример. Рассмотрим табл. 14.4. В ней пересекаются группировки посевных площадей по категориям хозяйств и группам сельскохозяйственных культур.

В двенадцати клетках таблицы над диагоналями приведены доли культур в итогах по категориям хозяйств, например, зерновые составляют 54,38% всей площади посева сельскохозяйственных предприятий. Под диагоналями приводятся доли категорий хозяйств в итогах площади каждой группы культур. Доля сельскохозяйственных предприятий в общей площади зерновых культур составила 90,56%. В итоговой строке

Таблица 14.4

Структура посевных площадей в России по категориям хозяйств и по группам культур, 1999 г. (в процентах)

Категория хозяйств	Группы сельскохозяйственных культур				Итого
	зерновые	технические	картофель и овощи	кормовые	
Сельскохозяйственные предприятия	54,38 / 90,56	8,12 / 82,89	0,64 / 11,90	36,98 / 95,67	100 / 87,88
Хозяйства населения	6,38 / 0,64	2,13 / 1,32	78,72 / 85,72	12,77 / 2,00	100 / 5,32
Крестьянские (фермерские) хозяйства	66,67 / 8,80	20,00 / 15,79	1,67 / 2,38	11,66 / 2,33	100 / 6,80
Итого	52,77 / 100	8,49 / 100	4,76 / 100	33,98 / 100	100

над диагоналями приведены доли итогов по культурам в общем итоге всех посевных площадей России. Под диагоналями — итог долей категорий хозяйств в общей площади данной культуры, т.е. 100%. В итоговой графе над диагоналями приводятся итоги долей сельскохозяйственных культур в общей площади данной категории хозяйств, т.е. 100%. Под диагоналями приводятся доли площади у данной категории хозяйств во всех посевных площадях в России. Кроме указанных четырех видов долей можно вычислить и пятый вид: доли площадей под данной культурой в данной категории хозяйств от общей посевной площади всех культур во всех категориях хозяйств. Каждая такая доля равна произведению доли над диагональю во внутренних клетках таблицы и доли под диагональю в итоговой графе или произведению доли под диагональю клетки таблицы и доли над диагональю в итоговой строке.

Например, доля площади зерновых в сельскохозяйственных предприятиях в общей посевной площади России равна произведению доли зерновых в сельскохозяйственных предприятиях на долю сельскохозяйственных предприятий в общей площади посевов в России: $54{,}38\% \cdot 87{,}88\% = 47{,}79\%$ (не забудем, что произведение процентов на проценты дает десятитысячные доли). Эта же доля может быть получена как произведение доли сельскохозяйственных предприятий в итоге посевов зерновых на долю итога зерновых культур в общей площади посевов в России: $90{,}56\% \cdot 52{,}77\% = 47{,}79\%$.

Итак, двухмерная пересекающаяся структура позволяет рассчитать пять видов структурных показателей (долей). При трех пересекающихся признаках группировки число разных видов структур достигает 19. В общем виде при n взаимопересекающихся признаках структура содержит $(n^3 - n^2 + 1)$ видов долей.

Конечно, вовсе не обязательно при каждом конкретном исследовании вычислять все эти показатели. Исходить следует из поставленной задачи, и вычислять те виды показателей структуры, которые для данной задачи имеют существенное значение. В отличие от анализа балансовой структуры, где две стороны баланса взаимосвязаны, при анализе структуры с пересекающимися независимыми признаками соотношения между долями, образованными по равным группировочным

признакам, смысла не имеют, как, например, соотношение доли технических культур в сельскохозяйственных предприятиях с долей картофеля в хозяйствах населения. Если же группировочные признаки, образующие пересекающиеся структуры, статистически связаны друг с другом, то анализ такой структуры позволяет измерять тесноту связи (см. гл. 9).

14.5. Сравнительный анализ структур

Сравнение структурных показателей по разным признакам может служить важным аналитическим приемом исследования. Рассмотрим данные табл. 14.5.

Сопоставление абсолютных величин родившихся и умерших не раскрывает различия в естественном движении населения по субъектам Российской Федерации: во всех субъектах число умерших больше числа родившихся. Различие раскрывает сравнение структурных показателей: в Москве, Московской области, Санкт-Петербурге, Нижегородской области доля умерших намного превышает долю родившихся; в Баш-

Таблица 14.5

Структура естественного движения населения по некоторым субъектам Российской Федерации в 2001 г.

Субъект	Родилось за год		Умерло за год		Отношение доли умерших к доле родившихся
	тыс. чел.	% к РФ	тыс. чел.	% к РФ	
Российская Федерация	1311,6	100	2254,9	100	1,00
Москва	76,0	5,8	134,9	6,0	1,03
Московская область	51,1	3,9	117,3	5,2	1,33
Санкт-Петербург	33,8	2,6	76,2	3,3	1,27
Краснодарский край	48,4	3,7	77,6	3,4	0,92
Республика Башкортостан	42,8	3,2	55,0	2,4	0,75
Республика Татарстан	35,9	2,7	50,1	2,2	0,81
Нижегородская область	28,2	2,1	66,8	3,0	1,43
Челябинская область	33,8	2,6	56,5	2,5	0,96

Источник. Российский статистический ежегодник. 2002. Стат. сборник. — М.: Госкомстат России, 2002. — С. 105—118.

кортостане и Татарстане, наоборот, больше доля родившихся. Построив показатель соотношения долей (последняя графа табл. 14.5), видим, что худшее положение сложилось в Нижегородской области, второе место снизу занимает Московская область, затем — Санкт-Петербург. Краснодарский край, Челябинская область и Москва находятся примерно на среднероссийском уровне. Лучшее положение из перечисленных регионов занимает Республика Башкортостан, чья доля среди умерших на 25%, или почти на целый пункт, ниже доли родившихся. Подчеркнем еще раз, что полученные новые показатели соотношения структур нетождественны ни по величине, ни по содержанию коэффициентам рождаемости и смертности — ведь и в Республике Башкортостан смертность превышала рождаемость. Соотношение долей содержат новую информацию — в этом их значение.

Аналогично можно сравнить доли регионов в сумме средств, перечисляемых ими в федеральный бюджет, с долей получаемых из него дотаций и субсидий, долю страны в территории суши с долей добываемых в стране полезных ископаемых.

Так, Россия, занимая 10% площади суши Земли, добывает 11,6% нефти, 28,1% природного газа, 13% каменного угля. Это говорит о том, что Россия является мировым донором энергоносителей (а также алмазов, апатита, калийных солей и других ископаемых). Еще один яркий пример сравнения структурных показателей: Москва, имеющая 6,8% населения России, по сумме активов банков и обороту финансовых средств занимает в России 50%, что говорит о ненормально высокой степени концентрации финансового капитала в столице.

При изучении распределения населения страны по душевому доходу (табл. 14.6) структурный анализ и сравнение структур позволяют раскрыть характер этого распределения, имеющий ключевое значение для понимания социальной структуры общества и социальной политики государства.

На основе данных табл. 14.6 можно сравнить структурные показатели населения, с одной стороны, и доходов — с другой. Например, 10% беднейшего населения региона, т.е. первая децильная группа, имеют лишь 2,4% всех доходов, а 10% наиболее обеспеченного населения, т.е. десятая децильная

Таблица 14.6

Распределение населения региона по децильным группам душевого дохода, 2000 г.

Номер группы i	Среднедушевой доход, руб./чел. H_i	Доля лиц, % $d.f._i$	Средний доход, руб./чел. x_i	Доля дохода, % dx_i	Плотность распределения $10:H_i$	$\Delta_x = x_i - x$	Δ_x^2	$d_{xi} - d_{fi}$, %	Кумулятивная доля доходов, % d_x
1	До 702	10	505	2,4	0,025	−1598	2,55	−7,6	2,4
2	702—853	10	778	3,7	0,067	−1325	1,76	−6,3	6,1
3	853—1055	10	968	4,6	0,050	−1135	1,29	−5,4	10,7
4	1055—1256	10	1157	5,5	0,049	−946	0,89	−4,5	16,2
5	1256—1509	10	1367	6,5	0,040	−736	0,54	−3,5	22,7
6	1509—1850	10	1683	8,0	0,029	−420	0,18	−2,0	30,7
7	1850—2318	10	2124	10,1	0,021	+21	0,00	0,1	40,8
8	2318—2794	10	2524	12,0	0,021	+421	0,18	2,0	52,8
9	2795—3652	10	3135	14,9	0,012	+1031	1,07	4,9	67,7
10	Более 3652	10	6794	32,3	0,000	+4691	22,01	22,3	100
	Итого	100	21034	100	—	—	30,47	—	350,1

группа, имеют 32,3% всех доходов, или в 13,4 раза больше, чем первая группа.

Несовпадение структурных показателей — долей населения с долями дохода свидетельствует о неравномерности распределения доходов. Совпадение долей говорило бы о полной уравнительности распределения доходов, чего, конечно, ни в одной стране или регионе не бывает. Чем более крупные доли населения сравниваются, тем меньше становится соотношение долей доходов. Так, доходы 20% наиболее обеспеченного населения (группы 9 и 10) только в 7,7 раза превышают доходы 20% наименее обеспеченного населения (группы 1 и 2), а доходы 50% более обеспеченных превышают 50% менее обеспеченных в 3,4 раза. Напротив, соотношение доходов 5% наиболее богатых к доходам 5% наиболее бедных будет больше, чем соотношение доходов 10% и 10% тех и других. К сожалению, Госкомстат не публикует более подробного распределения, чем по децильным группам. Для сравнения можно привести данные («Известия» от 11.04.2000 г.) о том, что в

Бразилии разница доходов 20% богатейших к доходам 20% беднейших жителей достигает 26 раз (65% доходов против 2,5%)! В той же газете от 16.04.2000 г. сообщается, что 63 беднейшие страны мира, составляющие 57% всего населения Земли, имели лишь 6% мировых доходов, а 16,25% населения Земли, живущего в наиболее богатых странах (США, Канада, Европейский союз и Япония), имели 80% всех доходов. Показатель неравномерности распределения доходов, т.е. частное от деления долей доходов на доли населения, составил:

$$\frac{80}{16{,}25} : \frac{6}{57} = 46{,}77 \text{ раза.}$$

На этом фоне неравномерность распределения доходов в России значительно меньше.

При анализе распределения полезно сравнить среднюю арифметическую величину душевого дохода $\bar{x}$ с медианой и модой распределения (см. гл. 5). По данным табл. 14.6 $\bar{x} = \Sigma(\bar{x}_i) : 10 = 2103$ руб. на человека в месяц. Медиана находится на границе 50% менее обеспеченных и 50% более обеспеченных, т.е. на границе пятого и шестого интервалов, Me = 1509 руб. Мода находится в интервале с наибольшей плотностью распределения и равна:

$$\text{Mo} = 702 + \frac{(0{,}067 - 0{,}025) \cdot 150}{(0{,}067 - 0{,}025) + (0{,}067 - 0{,}050)} = 807 \text{ руб.}$$

Медиана почти вдвое больше моды, а средняя — в 2,6 раза. Сильное отличие между структурными средними и арифметической средней свидетельствует о сильной асимметрии распределения, его резком отличии от нормального, т.е. от распределения по закону Гаусса—Лапласа.

Специальными показателями степени неравномерности распределения служат коэффициент вариации, коэффициент Лоренца и коэффициент Джини.

По данным табл. 14.6 среднее квадратическое отклонение доходов от средней арифметической величины

$$\sigma = \sqrt{\frac{\sum_{1}^{10}(x_j - \bar{x})^2}{10}} = 1745 \text{ руб.;}$$

коэффициент вариации $v = \sigma : x = 1747 : 2103 = 0{,}83$, или 83%.

Как известно, коэффициент вариации не имеет верхней границы и может превышать 100%. Коэффициенты Лоренца и Джини принимают значения в границах от нуля до единицы и измеряют неравномерность распределения. Коэффициент Лоренца основан на прямом сравнении долей групп по числу единиц совокупности и долей по объему признака.

По данным табл. 14.6 коэффициент Лоренца составил:

$$L = \frac{\sum_{1}^{10}|dx_i - df_i|}{2}; \qquad L = \frac{58{,}6\%}{2} = 29{,}3\%,$$

где dx_i — доля i-й группы в объеме признака x;
df_i — доля i-й группы в числе единиц.

Знаменатель коэффициента — это максимальная величина модуля.

Коэффициент Джини основан на более сложной методике сравнения долей:

$$G = 1 - 2\sum_{1}^{10} df_j d'x_j + \sum_{1}^{n} df_j dx_j, \qquad (14.3)$$

при равночастотном распределении df_i — const = 0,1 формула упрощается:

$$G = 1 - 0{,}2\sum d'x_j + 0{,}1 \cdot 1.$$

По данным табл. 14.6

$$G = 1 - 0{,}2 \cdot 3{,}501 + 0{,}1 =$$

$$= 0{,}3998, \text{ или } \approx 40\%. \qquad (14.4)$$

По кумулятивным долям населения и доходов можно построить кривую Лоренца (рис. 14.2). На графике по оси абсцисс показаны кумулятивные доли населения, а по оси ординат — кумулятивные доли доходов. Соединив десять точек ломаной линией или плавной кривой, получим график факти-

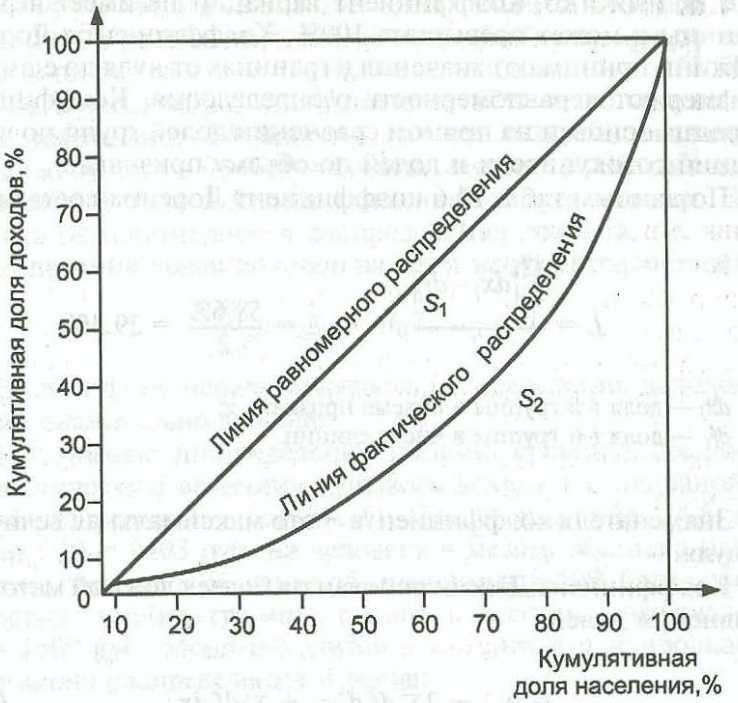


Рис. 14.2. Диаграмма Лоренца

ческого распределения, а проведя диагональ от начала координат до точки 100% по обеим осям, получим график равномерного распределения, которое означает, что 10% населения имели 10% всех доходов, 20% населения — $1/5$ доходов, 50% населения — половину доходов и т.п. Чем больше отстоит фактическая линия от диагонали, тем сильнее неравномерность распределения.

Степень неравномерности можно измерять и графически, как отношение площади S_1 между диагональю и кривой фактического распределения к сумме площадей $S_1 + S_2$ (технически это легко сделать, если график построен на бумаге в мелкую клетку. Тогда нужно просто сравнить число клеток в S_1 и S_2).

14.6. Показатели концентрации, специализации, монополизации. Многомерная структура

Методы и показатели анализа структуры используются при изучении таких важных экономических процессов, как концентрация производства, специализация предприятий или отраслей, диверсификация капитала, степень монополизации рынка и др.

В гл. 5 рассмотрены показатели специализации предприятий региона, зоны, основанные на измерении вариации объемов производства или долей предприятий, а также отношении фактических мер вариации к предельно возможным при данной численности совокупности. В подразд. 14.5 рассмотрены показатели концентрации объема признака, основанные на неравномерности его распределения между единицами совокупности. Но эти характеристики не являются исчерпывающими. Даже равномерное распределение производства, скажем, автомобилей в стране, где всего три предприятия, производящих по 33,3% всего выпуска автомашин, говорит о высокой степени концентрации в данной отрасли и вероятности его монополизации для устранения конкуренции и получения монопольной прибыли. Следовательно, показатель концентрации должен учитывать две величины: численность совокупности и степень неравномерности распределения признака между ее единицами. Рассмотрим методику конструирования показателя по заданным свойствам.

Проще построить показатель, учитывающий численность совокупности и быстроубывающий, как убывает степень концентрации и вероятности монополизации, с ростом числа производителей n. Можно эту составляющую желаемого показателя представить, например, как величину, обратную числу единиц совокупности, т.е. $1 : n$. При одном предприятии имеем абсолютный максимум, равный 1; при $n = 2$, $n = 3$, $n = 4$ доля все еще довольно значительна, но при большом n (большом числе производителей товара или услуг) эта составляющая уже становится несущественно малой, и существенное значение приобретает вторая составляющая — степень неравномерности распределения объема признака между единицами совокупности. Для того чтобы построить показатель, рассмотрим, как зависит от степени неравномерности распреде-

ления признака сумма накопленных долей объема признака $\sum_{j=1}^{n} d'_{xj}$ при условии, что изучаемая совокупность проранжирована в порядке нарастания долей объема признака.

При этом условии минимальная сумма накопленных долей будет в том случае, когда доли всех единиц совокупности, кроме последней, равны нулю и доля последней («монополиста») равна единице. Сумма накопленных долей тоже равна единице. Итак, $\sum_{j=1}^{n} d_{xj_{\min}} = 1$. Найдем теперь выражение для максимума этой суммы, которая согласно условию ранжирования образуется при строгом равенстве всех долей, каждая из которых будет равна $1 : n$.

Нарастающие доли будут $1 : n$, $2 : n$, $3 : n$, и т.д. до $n : n$, а их сумма, как сумма членов арифметической прогрессии, выражается как: $(1 : n)(1 + 2 + \ldots + n) = (1 : n)(n^2 + n) : 2 = (n + 1) : 2$. Чем дальше отстоит фактическая сумма накопленных долей от максимальной величины, тем сильнее неравномерность распределения. Следовательно, в числителе должны стоять величины: $(n + 1) : 2 - \sum_{j=1}^{n} d'_{xj}$. Для того чтобы измерить степень отклонения от равномерности распределения, нужно сравнить меру фактической неравномерности с максимально возможной, равной разности между максимально возможной суммой накопленных долей и минимальной их суммой, равной единице. Следовательно, знаменатель должен иметь вид: $[(n + 1) : 2] - 1 = (n + 1 - 2) : 2 = (n + 1) : 2$. Итак, показатель степени концентрации за счет неравномерности распределения имеет формулу:

$$\frac{\frac{n+1}{2} - \sum_{j=1}^{n} d'_{xj}}{\frac{n+1}{2}} = \frac{n+1-2\sum_{j=1}^{n} d'_{xj}}{n+1}. \qquad (14.5)$$

Теперь объединяем обе составляющие и получаем показатель степени концентрации объема признака в совокупности, состоящей из n единиц, проранжированных в порядке возрастания объема признака или доли данной единицы в общем объеме признака в совокупности. Обозначим его K:

$$K = \frac{1}{n} + \frac{n+1-2\sum_{j=1}^{n} d'_{xj}}{n+1}. \qquad (14.6)$$

Можно провести преобразование этой формулы, но, по нашему мнению, лучше сохранить выражения обеих составляющих частей, чтобы их разная природа оставалась явной для пользователя. Остается выяснить свойства предлагаемого показателя концентрации и меры возможности монополизации рынка. При единственном монополисте: $n = 1$, первое слагаемое будет равно единице, второе — нулю. В итоге весь коэффициент равен единице. При $n = 2$ и равномерном распределении объема признака

$$K = \frac{1}{2} + \frac{2+1-2\cdot(0{,}5+1)}{2+1} = 0{,}5.$$

При сосредоточении всего объема признака во втором предприятии

$$K = \frac{1}{2} + \frac{2+1-2\cdot(0+1)}{2+1} = 0{,}83.$$

Эта величина показателя K максимальна из возможных для $n = 2$. При росте n первое слагаемое уменьшается и при $n \to \infty$ стремится к нулю. Второе слагаемое при концентрации всего производства у одного предприятия всегда остается равным единице, значит, при абсолютной концентрации $K \to 1$ при $n \to \infty$. При полной равномерности, когда второе слагаемое равно нулю, $K \to 0$ при $n \to \infty$, как и должно быть логически. При реальных значениях распределений объема признака между единицами совокупности получаем промежуточные значения между 1 и $(1 : n)$.

Пример. Рассмотрим степень концентрации активов среди десяти крупных банков Санкт-Петербурга на 01.10.2000 г. по

данным Ассоциации коммерческих банков Санкт-Петербурга на примере табл. 14.7.

Таблица 14.7

Активы банков Санкт-Петербурга

Наименование банка	Сумма, млн руб.	Доля, % d_{xi}	Накопленная доля, % d'_{xi}
Банк «Санкт-Петербург»	3900	3,43	3,43
Балтийский банк	7100	3,61	7,04
Инкасбанк	4200	3,70	10,74
Балтонэким банк	5000	4,40	15,14
Петровский народный банк	7100	6,25	21,39
Креди Лионэ Русбанк	8600	7,57	28,96
ВНП Дрезднер банк, Россия	16200	14,26	43,22
Санкт-Петербургское отделение «Менатеп»	18000	15,84	59,06
Промстройбанк	18500	16,29	75,35
Санкт-Петербургское отделение Сбербанка России	28000	24,65	100,00
Итого	113600	100	364,33

Источник. «Банк» — приложение к газете «Известия» № 21.12.2000.

Показатель степени концентрации активов среди этих банков составил:

$$K = \frac{1}{10} + \frac{10 + 1 - 2 \cdot 3{,}6433}{10 + 1} = 0{,}3376.$$

Это означает, что концентрация активов банков составляет 33,76%, около 1/3 предельной величины концентрации. От монополии это весьма далеко, но степень концентрации активов значительная. Поскольку показатель K есть доля от предельной величины, проверку нулевой гипотезы проведем по формуле ошибки доли (см. гл. 7). Эта ошибка m_p составит:

$$m_p = \frac{\sqrt{k(1-k)}}{\sqrt{10}} = \frac{\sqrt{0{,}34 \cdot 0{,}66}}{\sqrt{10}} = 0{,}1499.$$

Критерий t-Стьюдента для нулевой гипотезы t равен: $\frac{0{,}3376}{0{,}1499} = 2{,}25$, что совпадает с табличным значением при 9

степенях свободы и значимости 0,05. Нулевую гипотезу можно отклонить, наличие концентрации активов банков неслучайно. Из последнего замечания следует, что при экономической оценке величины концентрации и возможностей монополизации нельзя принимать в расчет только величину какого-то одного показателя, но надо проверить, насколько однородным является сам объемный признак, распределение которого изучается.

14.7. Абсолютные и относительные показатели изменения структуры

Об особенностях измерения динамики относительных величин, в том числе и долей, было сказано в подразд. 12.3. Здесь излагаются показатели, характеризующие не изменение отдельной доли, а изменение структуры в целом, т.е. «структурный сдвиг». Нередко под этим понятием понимают хорошо и давно известные индексы влияния изменения структуры на среднюю величину относительного показателя, например показателя эффективности: производительности труда, себестоимости продукции, урожайности, рентабельности и т.п. Эти индексы измеряют не величину самого изменения структуры, а его влияние (см. гл. 11).

Обратимся к примеру (табл. 14.8). Эти данные свидетельствуют о существенном изменении долей ВВП, использованных на разные цели. Обобщающим абсолютным показателем изменения структуры может служить сумма модулей абсолютных изменений долей, выраженная в процентных пунктах:

$$Ad = \sum_{j=1}^{k} |d_{ij} - d_{0j}|. \qquad (14.7)$$

В 1999 г. по сравнению с 1992 г. это абсолютное изменение, обозначенное Ad, составило 41,2 процентного пункта.

Расчет среднего абсолютного изменения, приходящегося на одну долю (группу, единицу совокупности), не дает никакой добавочной информации, ибо отношение среднего изменения к величине средней доли тождественно суммарному изменению в отношении к сумме долей, равной 1. Зато очень важно определить, насколько сильно произошедшее измене-

Таблица 14.8

Изменение структуры использования ВВП России

Направления использования ВВП	Доля, %		$\lvert d_{1i} - d_{0i}\rvert$	$(d_{1i} - d_{0i})^2$	$\frac{\lvert d_{1i} - d_{0i}\rvert}{d_{0i}}$
	1992 г. d_{0i}	1999 г. d_{1i}			
Потребление домохозяйств	33,7	50,4	16,7	278,89	0,496
Потребление государственных учреждений и коммерческих организаций	16,2	18,2	2,0	4,00	0,123
Валовое наполнение	35,7	15,1	20,6	424,36	0,577
Сальдо экспорта-импорта	14,4	16,3	1,9	3,61	0,132
Итого	100	100	41,2	710,86	1,328

Источник. Россия в цифрах. Краткий статистический сборник. — М.: Госкомстат России, 2000. — С. 151.

ние структуры в сравнении с предельно возможной величиной суммы модулей. Логически ясно, что максимальная сумма модулей изменения долей равна 2. Например, была одна доля в пределе, равная нулю, другая, равная единице, а в следующем периоде наоборот. Сумма модулей разности долей равна 2. Теперь можно построить показатель степени интенсивности абсолютного структурного сдвига *KAd*.

$$KAd = \frac{\sum_{j=1}^{k} |d_{1j} - d_{0j}|}{2}. \qquad (14.8)$$

По данным табл. 14.8, $KAd = \frac{41,2}{2} = 20,6$ процентного пункта. Изменение структуры использования ВВП России за 7 лет реформ на 20,6 процентного пункта (более чем на 1/5 предельно возможного) следует признать очень значительным.

Для того чтобы избежать взаимопогашения разных по знаку изменений долей, вместо модулей можно применить квадраты и получить квадратическую меру абсолютного структурного сдвига в форме квадратического изменения долей:

$$\sigma_d = \sqrt{\frac{\sum_{i=1}^{k} (d_{ij} - d_{0j})^2}{k}}. \qquad (14.9)$$

По данным табл. 14.8,

$$\sigma_d = \sqrt{\frac{710,86}{4}} = 13,33 \text{ процентного пункта.}$$

В данном случае все доли изменились почти на одинаковое число пунктов, поэтому средняя квадратическая величина почти равна арифметической средней: 31,8 : 4 = 7,95. При резко различных изменениях долей квадратическое изменение ближе к наибольшему из изменений, чем арифметическая средняя. Предельная величина суммы квадратов изменения долей также равна 2, как и сумма модулей изменений долей, так как $(1^2) + (-1^2) = 2$. Для четырех долей максимальное значение $\sigma_d = \sqrt{2} : 4 = 0,71$. Фактическое значение составило: 0,1333 : 0,71 = 0,188, или 18,8% предельно возможного.

Абсолютные показатели изменения долей не учитывают величины долей базисного периода, т.е. считается, что изменение доли на 10 процентных пунктов не показывает, была ли доля до этого равна 2% или 50%. Такой подход недостаточен. Ведь первая из долей при увеличении на 10 процентных пунктов возросла в 6 раз, а вторая — только на 1/5. Очевидно, изменение структуры следует охарактеризовать и относительным показателем, измеряющим среднее относительное изменение долей. Рассмотрим построение этого показателя. Средний темп изменения долей, взвешенный по величине базисных долей, тождественно равен 1.

$$\frac{\sum_{j=1}^{k} \frac{d_{1j}}{d_{0j}} \cdot d_{0j}}{\sum_{j=1}^{k} d_{0j}} = \frac{\sum_{j=1}^{k} d_{1j}}{\sum_{j=1}^{k} d_{0j}} = \frac{1}{1}.$$

Невзвешенный средний темп изменения при разных долях не обязательно равен 1, но из-за взаимопогашения темпов, больших 1, и темпов, меньших 1, близок к 1, и ничто не говорит о мере изменения структуры. Наиболее информативным оказывается среднее относительное линейное изменение (темп прироста) по модулю:

$$I_d = \frac{\sum_{j=1}^{k} \left| \frac{d_{ij} - d_{0j}}{d_{0j}} \right|}{k}. \qquad (14.10)$$

По данным табл. 14.8, эта величина составляет:

$$I_d = \frac{1{,}328}{4} = 0{,}332, \text{ или } 33{,}2\% \text{ (а не пункта).}$$

Этот показатель означает, что при изменении структуры использования ВВП России произошел в среднем 33%-ный сдвиг — изменение роли статей в итоге. Величина I_d предела не имеет, так как малая доля может возрасти в бесконечно большое число раз. Использовать необходимо лишь простую среднюю из относительных изменений долей, так как средняя величина, взвешенная по базисным долям, как легко можно убедиться, всегда равна ранее рассмотренному абсолютному изменению *Ad*.

К. Гатевым, С. В. Курышевой, Т. Н. Агаповой предложен еще ряд показателей относительного изменения структуры, о которых желающие расширить свои знания могут прочитать в указанной в конце главы литературе, а также в подразд. 6.2. Отметим особо лишь показатель К. Гатева:

$$K_\Gamma = \sqrt{\frac{\sum_{(i)} (d_{i1} - d_{i0})^2}{\sum_{(i)} d_{i1}^2 + \sum_{(i)} d_{i0}^2}}.$$

Этот интегральный коэффициент структурных сдвигов изменяется в пределах [0, 1]; чем ближе к 1, тем сильнее изменение структуры.

14.8. Ранговые показатели изменения структуры

Изменения структуры не сводятся к возрастанию и уменьшению долей элементов этой структуры. В ряде практических задач особую роль играют ранги долей. Представим себе, что в каком-то комитете, на конференции, в Государственной Думе РФ и т.д. обсуждался законопроект и по мере внесения в него поправок проводились три голосования, результаты которых представлены в табл. 14.9.

При втором голосовании в сравнении с первым произошло существенное изменение структуры вотумов: абсолютное изменение (по модулю): $A_{d2/1} = 17 + 3 + 14 = 34$ процентным пунктам, среднее изменение — по 11,33 пункта на элемент. Абсолютный сдвиг при третьем голосовании в сравнении со вторым намного скромнее: $A_{d3/2} = 6 + 5 + 1 = 12$, или 4 пунктам на элемент структуры. Однако качественное различие структур второго и первого голосований не принципиально. И в первом, и во втором голосовании законопроект не принят, а в третьем он одобрен. Это качественное различие проявилось в изменении рангов вотумов. Аналогичную ситуацию имеем в ряде других явлений. Так, в результате экзаменационной сессии ранг («место», занятое группой) может быть гораздо важнее (скажем, группы, занявшие I и II места, награждаются путевкой, ценным призом), чем величина различия в долях отличников, хорошистов, троечников и двоечников. Изменение рангов статей платежного баланса страны, рангов статей в структуре ВВП может иметь гораздо большее экономическое значение, чем даже значительный абсолютный структурный сдвиг без изменения рангов.

Таблица 14.9

Результаты голосования по законопроекту

Вид вотума	Результаты голосования, %			Ранги вотума		
	I	II	III	I	II	III
За принятие	29	46	52	2	2	1
Против	54	51	46	1	1	2
Воздержались	11	3	2	3	3	3
Итого	100	100	100	—	—	—

На основе изменения рангов долей можно построить два показателя.

1. *Линейный коэффициент изменения рангов долей.* Обозначим его *KR*. Он представляет собой отношение фактической суммы модулей изменения рангов к предельно возможной сумме модулей при n элементах структуры, равной $(n^2 : 2)$ для четного и $(n^2 - 1) : 2$ для нечетного n:

$$KR = \frac{\sum_{i=1}^{n} |R_{1i} - R_{0i}|}{n^2 : 2} \quad \text{или} \quad KR = \frac{\sum_{i=1}^{n} |R_{1i} - R_{0i}|}{(n^2 - 1) : 2}. \qquad (14.11)$$

По данным табл. 14.11 этот коэффициент составил:

$$KR_{3/2} = \frac{1+1}{(3^2-1):2} = 0{,}5, \text{ или } 50\%.$$

Изменение рангов на 50% максимального, конечно, является существенным преобразованием структуры. Если подсчитать по ней ранги долей по данным табл. 14.10 получим: $KR_{1/0} = \frac{2}{4^2 : 2} = 0{,}25$, или 25% максимального, что также следует признать значительным изменением. О социально-экономическом значении этого изменения («хорошо» или «плохо») можно спорить, поскольку сокращение доли накопления, да еще при абсолютном снижении всего объема ВВП подрывает перспективы роста экономики.

2. *Квадратический коэффициент изменения рангов долей (KRK).* Для его построения используем известный коэффициент корреляции рангов Спирмена (см. гл. 11).

При полном совпадении рангов долей в базисном и текущем периодах коэффициент Спирмена равен +1. При максимальном изменении рангов (первый становится последним, порядок рангов «переворачивается») коэффициент Спирмена составит −1. Следовательно, максимальное значение изменения коэффициента Спирмена равно 2. Для того чтобы построить показатель степени интенсивности изменения рангов элементов структуры, следует отклонение фактического коэффициента Спирмена от единицы разделить на 2. Получим формулу *KRK*:

$$KRK = \frac{1 - \left(1 - \frac{6\sum_{i=1}^{n}(R_{1i} - R_{0i})^2}{n^3 - n}\right)}{2} = \frac{3\sum_{i=1}^{n}(R_{1i} - R_{0i})^2}{n^3 - n}, \qquad (14.12)$$

где R_{1i} и R_{0i} — ранги долей элементов структуры в базисном и отчетном периодах.

Измерим с помощью этого показателя сдвиг в ранжировании банков Санкт-Петербурга по сумме активов, используя данные табл. 14.7 и тех из входящих в нее банков, которые существовали и в 1995 г., и получим табл. 14.10.

$KRK = \frac{3 \cdot 62}{7^3 - 7} = \frac{186}{336} = 0{,}554$, что свидетельствует об очень существенном изменении рангов санкт-петербургских крупных банков за 1995—2000 гг. В основном это изменение связано с финансовым кризисом осени 1998 г.

Приведенные примеры показывают, что при анализе изменения структуры следует применить не какой-то один показатель, а всю их систему, так как каждый показатель отражает, измеряет особый аспект структурного сдвига. Разные

Таблица 14.10

Ранги банков Санкт-Петербурга по сумме активов

Наименование банка	Ранг		$R_1 - R_0$	$(R_1 - R_0)^2$
	1995 г. R_0	2000 г. R_1		
Санкт-Петербургское отделение Сбербанка России	3	1	−2	4
Промстройбанк	2	2	0	0
Банк «Санкт-Петербург»	1	7	6	36
Балтийский банк	5	6	1	1
Петровский народный банк	4	5	1	1
Креди Лионэ Русбанк	6	4	−2	4
ВНП Дрезднер банк, Россия	7	3	−4	16
Итого	—	—	—	62

показатели изменения структуры связаны между собой не жесткой связью, а связью статистической, в среднем — прямой зависимостью, но в конкретных процессах изменения структуры разные показатели могут сильно расходиться и даже изменяться в разных направлениях.

Изменение структуры сложных систем включает не только изменение состава и долей материальных элементов структуры, но также изменение структуры связей между этими элементами.

Изменение структуры коэффициента детерминации, состоящего, как показано в гл. 9, из суммы квадратов бета-коэффициентов и системного эффекта η_0, можно измерить показателем:

$$KAD = \frac{\sum_{i=1}^{k} |\beta_{1i}^2 - \beta_{0i}^2| + |\eta_1 - \eta_0|}{R_1^2 + R_2^2}, \qquad (14.13)$$

где β_{1i} и β_{0i} — стандартизованные коэффициенты регрессии в отчетном и базисном периодах;

η_1 и η_0 — индексы корреляции;

R_1 и R_0 — коэффициенты множественной корреляции.

Применение показателя *KAD* возможно при постоянстве системы взаимосвязанных признаков в отчетном и базисном периодах.

РЕЗЮМЕ

Понятие «структура совокупности» является базовым и используется в решении разнообразных задач.

Структура — форма организации системы, состоящей из отдельных элементов и связей между ними.

Изучение структуры и структурных изменений зависит от характера структуры. Различают иерархическую (древовидную) и неиерархическую структуры, балансовую, многомерную структуры с пересекающимися признаками.

В изучении динамики структуры или степени соответствия структур разных территориальных объектов особый интерес представляет случай, когда составные элементы структуры неравновелики и нужно определить изменения за счет крупных элементов с малой динамикой и мелких элементов с большой динамикой.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Агапова Т. Н.* Методы статистического изучения структуры сложных систем и ее изменения. — М.: Финансы и статистика, 1996.
2. *Гатев К.* Статистическая оценка различий между структурами // Теоретические и методологические проблемы статистики. — М.: Статистика, 1979.
3. *Елисеева И. И., Рукавишников В. О.* Группировка, корреляция, распознавание образов. — М.: Статистика, 1977.
4. *Казинец Л. С.* Измерение структурных сдвигов в экономике. — М.: Экономика, 1969.
5. *Казинец Л. С.* Темпы роста и структурные сдвиги в экономике. — М.: Экономика, 1981.
6. *Курышева С. В.* Статистический анализ содержания труда рабочих. — Красноярск: Изд-во КГУ, 1990.
7. *Миркин Б. Г.* Анализ качественных признаков и структур. — М.: Статистика, 1980.

ПРИЛОЖЕНИЯ

1. Статистико-математические таблицы

Таблица П.1

Значение интеграла вероятностей $F(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt$

t	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0718
0,1	0797	0876	0955	1034	1114	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3752
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4909	4971	5035	5098
0,7	5161	5223	5285	5346	5467	5497	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6626	6679	6729	6778
1,0	6817	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8030
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8788	8812.	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9089
1,7	9108	9127	9146	9164	9182	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9425	9438	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9608	9615	9625	9634
2,1	9643	9652	9660	9669	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9755	9762	9768	9774	9780
2,3	9785	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9840	9845	9849	9853	9857	9861	9866	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9909	9912	9915	9917	9920	9924	9926	9927	9929
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947

Продолжение

t	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
2,8	9949	9950	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	99730	99739	99747	99755	99763	99771	99779	99786	99793	99800
3,1	99807	99813	99819	99825	99831	99837	99842	99847	99853	99858
3,2	99863	99867	99872	99876	99880	99884	99888	99892	99896	99900
3,3	99903	3,6	99911	3,9	999904	4,4	9999892	5,0		99999943
3,4	99933	3,7	99937	4,0	999937	4,6	9999957	5,0		99999996
3,5	99953	3,8	99957	4,2	999973	4,8	9999984	6,0		999999998

Таблица П.2

Значение t-критерия Стьюдента при уровне значимости 0 10; 0 05; 0,01

Число степеней свободы, *d.f.*	p			Число степеней свободы, *d.f.*	p		
	0,10	0,05	0,01		0,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,6041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6449	1,9600	2,5758

Значение *F*-критерия Фишера

Таблица П.3

при уровне значимости 0,05

$d.f._2$	$d.f._1$															
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	30	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	250	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41	19,42	19,43	19,44	19,46	19,50
3	10,13	9,55	9,28	9,19	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74	8,71	8,69	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91	5,87	5,84	5,80	5,74	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68	4,64	4,60	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,96	3,92	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57	3,52	3,49	3,44	3,38	3,23
8	5,32	4,46	4,07	3,84	3,69	3,.58	3,50	3,44	3,39	3,34	3,31	3,28	3,23	3,20	3,15	3,08	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07	3,02	2,98	2,93	2,86	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	.2,94	2,91	2,86	2,82	2,77	2,70	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79	2,74	2,70	2,65	2,57	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69	2,64	2,60	2,54	2,46	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60	2,55	2,51	2,46	2,38	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53	2,48	2,44	2,39	2,31	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48	2,43	2,39	2,33	2,25	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42	2,37	2,33	2,28	2,20	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38	2,33	2,29	2,23	2,15	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,29	2,25	2,19	2,11	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,34	2,31	2,26	2,21	2,15	2,07	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31	2,28	2,23	2,18	2,12	2,04	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,20	2,15	2,09	2,00	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,26	2,23	2,18	2,13	2,07	1,98	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28	2,24	2,20	2,14	2,10	2,04	1,96	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22	2,18	2,13	2,09	2,02	1,94	1,73
25	4,24	3,88	2,99	2,76	2,60	2,49	2,41	2,34	2,26	2,24	2,20	2,16	2,11	2,06	2,00	1,92	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,10	2,05	1,99	1,90	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	2,20	2,16	2,13	2,08	2,03	1,97	1,88	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,15	2,12	2,06	2,02	1,96	1,87	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,05	2,00	1,94	1,85	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12	2,09	2,04	1,99	1,93	1,84	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04	2,00	1,95	1,90	1,84	1,74	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98	1,95	1,90	1,85	1,78	1,69	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,86	1,81	1,75	1,65	1,39
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88	1,85	1,79	1,75	1,68	1,57	1,28
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79	1,75	1,69	1,64	1,57	1,46	1,00

Примечание. $d.f._1$ — число степеней свободы для большей дисперсии; $d.f._2$ — число степеней свободы для меньшей дисперсии.

Таблица П.4

Значение χ^2-критерия Пирсона при уровне значимости 0,10; 0,05; 0,01

d.f.	0,10	0,05	0,01	*d.f.*	0,10	0,05	0,01
1	2,71	3,84	6,63	21	29,62	32,67	38,93
2	4,61	5,99	9,21	22	30,81	33,92	40,29
3	6,25	7,81	11,34	23	32,01	35,17	41,64
4	7,78	9,49	13,28	24	33,20	36,42	42,98
5	9,24	11,07	15,09	25	34,38	37,65	44,31
6	10,64	12,59	16,81	26	35,56	38,89	45,64
7	12,02	14,07	18,48	27	36,74	40,11	46,96
8	13,36	15,51	20,09	28	37,92	41,34	48,28
9	14,68	16,92	21,67	29	39,09	42,56	49,59
10	50,99	18,31	23,21	30	40,26	43,77	50,89
11	17,28	19,68	24,72	40	51,80	55,76	63,69
12	18,55	21,03	26,22	50	63,17	67,50	76,15
13	19,81	22,36	27,69	60	74,40	79,08	88,38
14	21,06	23,68	29,14	70	85,53	90,53	100,42
15	22,31	25,00	30,58	80	96,58	101,88	112,33
16	23,54	26,30	32,00	90	107,56	113,14	124,12
17	24,77	27,59	33,41	100	118,50	124,34	135,81
18	25,99	28,87	34,81				
19	27,20	30,14	36,19				
20	28,41	31,14	37,57				

Таблица П.5

Критические значения коэффициентов корреляции для уровней значимости 0,05; 0,01

d.f.	$\alpha = 0,05$	$\alpha = 0,01$	*d.f.*	$\alpha = 0,05$	$\alpha = 0,01$
1	0,996917	0,9998766	17	0,4555	0,5751
2	0,950000	0,990000	18	0,4438	0,5614
3	0,8783	0,95873	19	0,4329	0,5487
4	0,8114	0,91720	20	0,4227	0,5368
5	0,7545	0,8745	25	0,3809	0,4869
6	0,07067	0,8343	30	0,3494	0,4487
7	0,6664	0,7977	35	0,3246	0,4182
8	0,6319	0,7646	40	0,3044	0,3932
9	0,6021	0,7348	45	0,2875	0,3721
10	0,5760	0,7079	50	0,2732	0,3541
11	0,5529	0,6835	60	0,2500	0,3248
12	0,5324	0,6614	70	0,2919	0,3017
13	0,5139	0,6411	80	0,2172	0,2830
14	0,4973	0,6226	90	0,2050	0,2673
15	0,4821	0,6055	100	0,1946	0,2540
16	0,4683	0,5897			

Для простой корреляции *d.f.* на 2 меньше, чем число пар вариантов; в случае частной корреляции необходимо также вычесть число исключаемых переменных.

Таблица П.6

Z*-преобразование. Значения величины *z* для значений *r

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1105	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
0,3	0,3095	0,3206	0,3317	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
0,4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,5	0,5493	0,5627	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,6	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467

Таблица П.7

Таблица случайных чисел

Ряд	Колонка							
	12345	67890	12345	67890	12345	67890	12345	67890
01	66194	28926	99547	16625	45515	67953	12108	57846
02	78240	43195	24837	32511	70880	22070	52622	61881
03	00833	88000	67299	68215	11274	55624	32991	17436
04	12111	86683	61270	58036	64192	90611	15145	01748
05	47189	99951	05755	03834	43782	90599	40282	51417
06	76396	72486	62423	27618	84184	78922	73561	52818
07	46409	17469	32483	09083	76175	19985	26309	91536
08	74626	22111	87286	46772	42243	68046	44250	42439
09	34450	81974	93723	49023	58432	67083	36876	93391
10	36327	72135	33005	28701	34710	49359	50693	89311
11	74185	77536	84825	09934	99103	09325	67389	45869
12	12296	41623	62873	37943	25584	09609	63360	47270
13	90822	60280	88925	99610	42772	60561	76873	04117
14	72121	79152	96591	90305	10189	79778	68016	13747
15	95268	41377	25684	08151	61816	58555	54305	86189
16	92603	09091	75884	93424	72586	88903	30061	14457
17	18813	90291	05275	01223	79607	95426	34900	09778
18	38840	26903	28624	67157	51986	42865	14508	49315
19	05959	33836	53758	16562	41081	38012	41230	20528
20	85141	21155	99212	32685	51403	31926	69813	58781
21	75047	59643	31074	38172	03718	32119	69506	67143
22	30752	95260	68032	62871	58781	34143	68790	69766
23	22986	82575	42187	62295	84295	30634	66562	31442
24	99439	86692	90348	66036	48399	73451	26698	39437
25	20389	93029	11881	71685	65452	89047	63669	02656
26	39249	05173	68256	36359	20250	68686	05947	09335
27	96777	33605	29481	20063	09398	01843	35139	61344
28	04860	32918	10798	50492	52655	33359	94713	28393
29	41613	42375	00403	03656	77580	87772	86877	57085
30	17930	00794	53836	53692	67135	98102	61912	11246
31	24649	31845	25736	75231	83808	98917	93829	99430
32	79899	34061	54308	59358	56462	58166	97302	86828
33	76801	49594	81002	30397	52728	15101	72070	33706

Продолжение

Ряд	Колонка							
	12345	67890	12345	67890	12345	67890	12345	67890
34	36239	63636	38140	65731	39788	06872	38971	53363
35	07392	64449	17886	63632	53995	17574	22247	62607
36	67133	04181	33874	98835	67453	59734	76381	63455
37	77759	31504	32832	70861	15152	29733	75371	39174
38	85992	72268	42920	20810	29361	51423	90306	73574
39	79553	75952	54116	65553	47139	60579	09165	85490
40	41101	17336	48951	53674	17880	45260	08575	49321
41	36191	17095	32123	91576	84221	78902	82010	30874
42	62329	63898	23268	74283	26091	68409	69704	82267
43	14751	13151	93115	01437	56945	89661	67680	79790
44	48462	59278	44185	29616	76537	19589	83139	28454
45	29435	88105	59651	44391	74588	55114	80834	85686
46	28340	29285	12965	14821	80425	16602	44653	70467
47	02167	58940	27149	80242	10587	79786	34959	75339
48	17864	00991	39557	54981	23588	81914	37609	13128
49	79675	80605	60059	35862	00254	36546	21545	78179
50	72335	82037	92003	34100	29879	46613	89720	13274
51	49280	88924	35779	00283	81163	07275	89863	02348
52	61870	41657	07468	08612	98083	97349	20775	45091
53	43898	65923	25078	86129	78496	97653	91550	08078
54	62993	93912	30454	84598	56095	20664	12872	64647
55	33850	58555	51438	85507	71865	79488	76783	31708
56	55336	71264	88472	04334	63919	36394	11095	92470
57	70543	29776	10087	10072	55980	64688	68239	20461
58	89382	93809	00796	95945	34101	81277	66090	88872
59	37818	72142	67140	50785	22380	16703	53362	44940
60	60430	22834	14130	96593	23298	56203	92671	15925
61	82975	66158	84731	19436	55790	69229	28661	13675
62	39087	71938	40355	54324	08401	26299	49420	59208
63	55700	24586	93247	32596	11865	63397	44251	43189
64	14756	23997	78643	75912	83832	32768	18928	57070
65	32166	53251	70654	92827	63491	04233	33825	69662
66	23236	73751	31888	81718	06546	83246	47651	04877
67	45794	26926	15130	82455	78305	55058	52551	47182

Продолжение

Ряд	Колонка							
	12345	67890	12345	67890	12345	67890	12345	67890
68	09893	20505	14225	68514	46427	56788	96297	78822
69	54382	74598	91499	14523	68479	27686	46162	83554
70	94750	89923	37089	20048	80336	94598	26940	36858
71	70297	34135	53140	33340	42050	82341	44104	82949
72	85157	47954	32979	26575	57600	40881	12250	73742
73	11100	02340	12860	74697	96644	89439	28707	25815
74	36871	50775	30592	57143	17381	68856	25853	35041
75	23913	48357	63308	16090	51690	54607	72407	55538
76	79348	36085	27973	65157	07456	22255	25626	57054
77	92074	54641	53673	54421	18130	60103	69593	49464
78	06873	21440	75593	41373	49502	17972	82578	16364
79	12478	37622	99659	31065	83613	69889	58869	29571
80	57175	55564	65411	42547	70457	03426	72937	83792
81	91616	11075	80103	07831	59309	13276	26710	73000
82	78025	73539	14621	39044	47450	03197	12787	47709
83	27587	67228	80145	10175	12822	86687	65530	49325
84	16690	20427	04251	64477	73709	73945	92396	68263
85	70183	58065	65489	31833	82093	16747	10386	59293
86	90730	35385	15679	99742	50866	78028	75573	67257
87	10934	93242	13431	24590	02770	48582	00906	58595
88	82462	30166	79613	47416	13389	80268	05085	96666
89	27463	10433	07606	16285	93699	60912	94532	95632
90	02979	52997	09079	92709	90110	47506	53693	49892
91	46888	69929	75233	52507	32097	37594	10067	67327
92	53638	83161	08289	12639	08141	12640	28437	09268
93	82433	61427	17239	89160	19666	08814	37841	12847
94	35766	31672	50082	22795	66948	65581	84393	15890
95	10853	42581	08792	13257	61973	24450	52351	16602
96	20341	27398	72906	63955	17276	10646	74692	48438
97	54458	90542	77563	51839	52901	53355	83281	19177
98	26337	66530	16687	35179	46560	00123	44546	79896
99	34314	23729	85264	05575	96855	23820	11091	79821
00	28603	10708	68933	34189	92166	15181	66628	58599

Таблица П.8

Значения функции Пуассона: $\frac{\lambda^m}{m!} e^{-\lambda}$

m \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,9048	0,8187	0,7408	0,6703	0,6055	0,5488	0,4966	0,4493	0,4066
1	0,0905	0,1638	0,2222	0,2681	0,3033	0,3293	0,3476	0,3596	0,3696
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647
3	0,0002	0,001	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494
4	—	—	0,0002	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111
5	—	—	—	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020
6	—	—	—	—	—	—	0,0001	0,0002	0,0003

m \ λ	1,0	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0
0	0,3679	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001
1	0,3679	0,2707	0,1494	0,0733	0,0337	0,0149	0,0064	0,0027	0,0011
2	0,1839	0,2707	0,2240	0,1465	0,0842	0,0446	0,0223	0,0107	0,0055
3	0,0313	0,1804	0,2240	0,1954	0,1404	0,0892	0,0521	0,0286	0,0150
4	0,0153	0,0902	0,1680	0,1954	0,1755	0,1339	0,0912	0,0572	0,0337
5	0,0081	0,0361	0,1008	0,1563	0,1755	0,1606	0,1277	0,0916	0,0607
6	0,0006	0,0120	0,0504	0,1042	0,1462	0,1606	0,1490	0,1221	0,0911
7	0,0001	0,0034	0,0216	0,0595	0,1044	0,1377	0,1490	0,1396	0,1318
8	—	0,0009	0,0081	0,0298	0,0656	0,1033	0,1304	0,1396	0,1318
9	—	0,0002	0,0027	0,0132	0,0363	0,0688	0,1014	0,1241	0,0318
10	—	—	0,0008	0,0053	0,0181	0,0413	0,0710	0,0993	0,1180
11	—	—	0,0002	0,0019	0,0082	0,0225	0,0452	0,0722	0,0970
12	—	—	0,0001	0,0006	0,0034	0,0113	0,0264	0,0481	0,0728
13	—	—	—	0,0002	0,0013	0,0052	0,0142	0,0296	0,0504
14	—	—	—	0,0001	0,0005	0,0022	0,0071	0,0169	0,0324
15	—	—	—	—	0,0002	0,0009	0,0033	0,0090	0,0194
16	—	—	—	—	—	0,0003	0,0014	0,0045	0,0109
17	—	—	—	—	—	0,0001	0,0006	0,0021	0,0058
18	—	—	—	—	—	—	0,0002	0,0009	0,0029
19	—	—	—	—	—	—	0,0001	0,0004	0,0014
20	—	—	—	—	—	—	—	0,0002	0,0006
21	—	—	—	—	—	—	—	0,0001	0,0003
22	—	—	—	—	—	—	—	—	0,0001

Таблица П.9

Критические значения *D*-критерия Колмогорова—Смирнова

Объем выборки	Уровень значимости, α				
	0,20	0,15	0,10	0,05	0,01
$n = 1$	0,900	0,925	0,950	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,708	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,510	0,565	0,669
6	0,410	0,436	0,470	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,360	0,388	0,432	0,514
10	0,322	0,342	0,368	0,410	0,490
11	0,307	0,326	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,450
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,250	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,20	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
Свыше 35	$\frac{1,07}{\sqrt{n}}$	$\frac{1,14}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Таблица П.10

Нижние и верхние значения критерия знаков Вилкоксона (*W*)

Двусторонняя проверка	α = 0,20	α = 0,10	α = 0,05	α = 0,02	α = 0,01
Односторонняя проверка	α = 0,10	α = 0,05	α = 0,025	α = 0,01	α = 0,005
$n = 4$	1,9	0,10	0,10	0,10	0,10
5	3,12	1,14	0,15	0,15	0,15
6	4,17	3,18	1,20	0,21	0,21
7	6,22	4,24	3,25	1,27	0,28
8	9,27	6,30	4,32	2,34	1,35
9	11,34	9,36	6,39	4,41	2,43
10	15,40	11,44	9,46	6,49	4,51
11	18,48	14,52	11,55	8,58	6,60
12	22,56	18,60	14,64	10,68	8,70
13	27,64	22,69	18,73	13,78	10,81
14	32,73	26,79	22,83	16,89	13,92
15	37,83	31,89	26,94	20,100	16,104
16	43,93	36,100	30,106	24,112	20,116
17	49,104	42,111	35,118	28,125	24,129
18	56,115	48,123	41,130	33,138	28,143
19	63,127	54,136	47,143	38,152	33,157
20	70,140	61,149	53,157	44,166	38,172

Таблица П.11

Нижние и верхние критические значения критерия суммы рангов Вилкоксона

α = 0,025 (односторонняя) или α = 0,05 (двусторонняя)																		
n_1 : 3			4		5		6		7		8		9		10			
n_2 : 3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33		
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44		
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56		
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70		
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83		
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98		
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114		
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131		
α= 0,05 (односторонняя), α = 0,10 (двусторонняя)																		
n_1 : 3			4		5		6		7		8		9		10			
n_2 : 3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31		
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42		
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54		
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67		
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80		
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95		
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111		
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127		

Примечание. $n_1 \leqslant n_2$.

Таблица П.12

Критические значения коэффициента ранговой корреляции Спирмена для правосторонней проверки при α

Объем выборки	0,10	0,05	0,025	0,01	0,005	0,001
n = 4	0,8000	0,8000				
5	0,7000	0,8000	0,9000	0,9000		
6	0,6000	0,7714	0,8286	0,8857	0,9429	
7	0,5357	0,6786	0,7450	0,8571	0,8929	0,9643
8	0,5000	0,6190	0,7143	0,8095	6,8571	6,9286
9	0,4667	0,5833	0,6833	0,7667	0,8167	0,9000
10	0,4424	0,5515	0,6364	0,7333	0,7818	0,8667
11	0,4182	0,5273	0,6091	0,7000	0,7455	0,8364
12	0,3986	0,4965	0,5804	0,6713	0,7273	0,8182
13	0,3791	0,4780	0,5549	0,6429	0,6978	0,7912
14	0,3626	0,4593	0,5341	0,6220	0,6747	0,7670
15	0,3500	0,4429	0,5179	0,6000	0,6536	0,7464
16	0,3382	0,4265	0,5000	0,5824	0,6324	0,7265
17	0,3260	0,4118	0,4853	0,5637	0,6152	0,7083
18	0,3148	0,3994	0,4716	0,5480	0,5975	0,6904
19	0,3070	0,3895	0,4579	0,5333	0,5825	0,6737
20	0,2977	0,3789	0,4451	0,5203	0,5684	0,6586
21	0,2909	0,3688	0,4351	0,5078	0,5545	0,6455
22	0,2829	0,3597	0,4241	0,4963	0,5426	0,6318
23	0,2767	0,3518	0,4150	0,4852	0,5306	0,6186
24	0,2704	0,3435	0,4061	0,4748	0,5200	0,6070
25	0,2646	0,3362	0,3977	0,4654	0,5100	0,5962
26	0,2588	0,3299	0,3894	0,4564	0,5002	0,5856
27	0,2540	0,3236	0,3822	0,4481	0,4915	0,5757
28	0,2490	0,3175	0,3749	0,4401	0,4828	0,5660
29	0,2443	0,3113	0,3685	0,4320	0,4744	0,5567
30	0,2400	0,3059	0,3620	0,4251	0,4665	0,5479

Таблица П.13

Величины $-p\log_2 p$

p	0	1	2	3	4	5	6	7	8	9
0,00	—	0,0100	0,0179	0,0251	0,0319	0,0382	0,0443	0,0501	0,0557	0,0612
0,01	0,0664	0,0716	0,0766	0,0845	0,0862	0,0909	0,0955	0,0999	0,1043	0,1086
0,02	0,1129	0,1170	0,1211	0,1252	0,1291	0,0330	0,1369	0,1407	0,1444	0,1481
0,03	0,1518	0,1554	0,1589	0,1624	0,1659	0,1693	0,1727	0,1760	0,1793	0,1825
0,04	0,1858	0,1889	0,1921	0,1952	0,1983	0,2013	0,2043	0,2073	0,2103	0,2132
0,05	0,2161	0,2190	0,2218	0,2246	0,2274	0,2301	0,2329	0,2356	0,2383	0,2409
0,06	0,2435	0,2461	0,2487	0,2513	0,2538	0,2563	0,2588	0,2613	0,2637	0,2661
0,07	0,2686	0,2709	0,2733	0,2756	0,2780	0,2803	0,2826	0,2848	0,2871	0,2893
0,08	0,2915	0,2937	0,2959	0,2980	0,3002	0,3023	0,3044	0,3065	0,3086	0,3106
0,09	0,3127	0,3147	0,3167	0,3187	0,3207	0,3226	0,3246	0,3265	0,3284	0,3303
0,10	0,3322	0,3341	0,3359	0,3378	0,3398	0,3414	0,3432	0,3450	0,3468	0,3485
0,11	0,3503	0,3520	0,3537	0,3555	0,3571	0,3588	0,3605	0,3622	0,3638	0,3654
0,12	0,3671	0,3687	0,3703	0,3719	0,3734	0,3750	0,3766	0,3781	0,3796	0,3811
0,13	0,3826	0,3841	0,3856	0,3871	0,3886	0,3900	0,3915	0,3929	0,3943	0,3957
0,14	0,3971	0,3985	0,3999	0,4012	0,4026	0,4040	0,4053	0,4066	0,4079	0,4092
0,15	0,4105	0,4118	0,4131	0,4144	0,4156	0,4169	0,4181	0,4194	0,4206	0,4218
0,16	0,4230	0,4242	0,4254	0,4266	0,4277	0,4289	0,4301	0,4312	0,4323	0,4335
0,17	0,4346	0,4357	0,4368	0,4379	0,4390	0,4400	0,4411	0,4422	0,4432	0,4443
0,18	0,4453	0,4463	0,4474	0,4484	0,4494	0,4504	0,4514	0,4523	0,4533	0,4543
0,19	0,4552	0,4562	0,4571	0,4581	0,4590	0,4599	0,4608	0,4617	0,4626	0,4635
0,20	0,4644	0,4653	0,4661	0,4670	0,4678	0,4687	0,4695	0,4704	0,4712	0,4720
0,21	0,4728	0,4736	0,4744	0,4752	0,4760	0,4768	0,4776	0,4783	0,4791	0,4798
0,22	0,4806	0,4813	0,4820	0,4828	0,4835	0,4842	0,4849	0,4856	0,4863	0,4870
0,23	0,4877	0,4883	0,4890	0,4897	0,4903	0,4910	0,4916	0,4923	0,4949	0,4935
0,24	0,4941	0,4947	0,4954	0,4960	0,4966	0,4971	0,4977	0,4983	0,4989	0,4994
0,25	0,5000	0,5006	0,5011	0,5016	0,5022	0,5027	0,5032	0,5038	0,5043	0,5048
0,26	0,5053	0,5058	0,5063	0,5068	0,5072	0,5077	0,5082	0,5087	0,5091	0,5096
0,27	0,5100	0,5105	0,5109	0,5113	0,5118	0,5122	0,5126	0,5130	0,5134	0,5138
0,28	0,5142	0,5146	0,5150	0,5154	0,5158	0,5161	0,5165	0,5169	0,5172	0,5176
0,29	0,5179	0,5182	0,5186	0,5189	0,5192	0,5196	0,5199	0,5202	0,5205	0,5208
0,30	0,5211	0,5214	0,5217	0,5220	0,5222	0,5225	0,5228	0,5230	0,5233	0,5235
0,31	0,5238	0,5240	0,5243	0,5245	0,5247	0,5250	0,5252	0,5254	0,5256	0,5258
0,32	0,5260	0,5262	0,5264	0,5266	0,5268	0,5270	0,5272	0,5273	0,5275	0,5277

Продолжение

p	0	1	2	3	4	5	6	7	8	9
0,33	0,5278	0,5280	0,5281	0,5283	0,5284	0,5286	0,5287	0,5288	0,5289	0,5290
0,34	0,5292	0,5293	0,5294	0,5295	0,5296	0,5297	0,5298	0,5299	0,5299	0,5300
0,35	0,5301	0,5302	0,5302	0,5303	0,5304	0,5304	0,5305	0,5305	0,5305	0,5306
0,36	0,5306	0,5306	0,5307	0,5307	0,5307	0,5307	0,5307	0,5307	0,5307	0,5307
0,37	0,5307	0,5307	0,5307	0,5307	0,5307	0,5306	0,5306	0,5306	0,5305	0,5305
0,38	0,5304	0,5304	0,5303	0,5303	0,5302	0,5302	0,5301	0,5300	0,5300	0,5299
0,39	0,5298	0,5297	0,5296	0,5295	0,5294	0,5293	0,5292	0,5291	0,5290	0,5289
0,40	0,5288	0,5286	0,5285	0,5284	0,5283	0,5281	0,5280	0,5278	0,5277	0,5275
0,41	0,5274	0,5272	0,5271	0,5269	0,5267	0,5266	0,5264	0,5262	0,5260	0,5258
0,42	0,5256	0,5255	0,5253	0,5251	0,5249	0,5246	0,5244	0,5242	0,5240	0,5238
0,43	0,5236	0,5233	0,5231	0,5229	0,5226	0,5224	0,5222	0,5219	0,5217	0,5214
0,44	0,5211	0,5209	0,5206	0,5204	0,5201	0,5198	0,5195	0,5193	0,5190	0,5187
0,45	0,5184	0,5181	0,5178	0,5175	0,5172	0,5169	0,5166	0,5163	0,5160	0,5157
0,46	0,5153	0,5150	0,5147	0,5144	0,5140	0,5137	0,5133	0,5130	0,5127	0,5123
0,47	0,5120	0,5116	0,5112	0,5109	0,5105	0,5102	0,5098	0,5094	0,5090	0,5087
0,48	0,5083	0,5079	0,5075	0,5071	0,5067	0,5063	0,5059	0,5055	0,5051	0,5047
0,49	0,5043	0,5039	0,5034	0,5030	0,5026	0,5022	0,5017	0,5013	0,5009	0,5004
0,50	0,5000	0,4996	0,4991	0,4987	0,4982	0,4978	0,4973	0,4968	0,4964	0,4959
0,51	0,4954	0,4950	0,4945	0,4940	0,4935	0,4930	0,4926	0,4921	0,4916	0,4911
0,52	0,4906	0,4901	0,4896	0,4891	0,4886	0,4880	0,4875	0,4870	0,4865	0,4860
0,53	0,4854	0,4849	0,4844	0,4839	0,4833	0,4828	0,4822	0,4817	0,4811	0,4806
0,54	0,4800	0,4795	0,4789	0,4784	0,4778	0,4772	0,4767	0,4761	0,4755	0,4750
0,55	0,4744	0,4738	0,4732	0,4726	0,4720	0,4714	0,4708	0,4702	0,4697	0,4691
0,56	0,4684	0,4678	0,4672	0,4666	0,4660	0,4654	0,4648	0,4641	0,4635	0,4629
0,57	0,4623	0,4616	0,4610	0,4603	0,4597	0,4591	0,4584	0,4578	0,4671	0,4565
0,58	0,4558	0,4551	0,4545	0,4538	0,4532	0,4525	0,4518	0,4512	0,4505	0,4498
0,59	0,4491	0,4484	0,4477	0,4471	0,4464	0,4457	0,4450	0,4443	0,4436	0,4429
0,60	0,4422	0,4415	0,4408	0,4401	0,4393	0,4386	0,4379	0,4372	0,4365	0,4357
0,61	0,4350	0,4343	0,4335	0,4328	0,4321	0,4313	0,4306	0,4298	0,4291	0,4283
0,62	0,4276	0,4268	0,4261	0,4253	0,4246	0,4238	0,4230	0,4223	0,4215	0,4207
0,63	0,4199	0,4192	0,4184	0,4176	0,4168	0,4160	0,4153	0,4145	0,4137	0,4129
0,64	0,4121	0,4113	0,4105	0,4097	0,4089	0,4080	0,4072	0,4064	0,4056	0,4048
0,65	0,4040	0,4032	0,4023	0,4015	0,4007	0,3998	0,3990	0,3982	0,3973	0,3965
0,66	0,3957	0,3948	0,3940	0,3931	0,3922	0,3914	0,3905	0,3897	0,3888	0,3880

Продолжение

p	0	1	2	3	4	5	6	7	8	9
0,67	0,3871	0,3862	0,3854	0,3845	0,3836	0,3828	0,3819	0,3810	0,3801	0,3792
0,68	0,3784	0,3775	0,3766	0,3757	0,3748	0,3739	0,3730	0,3721	0,3712	0,3703
0,69	0,3694	0,3685	0,3676	0,3666	0,3657	0,3648	0,3639	0,3630	0,3621	0,3611
0,70	0,3602	0,3593	0,3558	0,3574	0,3565	0,3555	0,3546	0,3536	0,3527	0,3518
0,71	0,3508	0,3499	0,3489	0,3480	0,3470	0,3461	0,3451	0,3441	0,3432	0,3422
0,72	0,3412	0,3403	0,3393	0,3383	0,3373	0,3364	0,3354	0,3344	0,3334	0,3324
0,73	0,3314	0,3304	0,3295	0,3285	0,3275	0,3265	0,3255	0,3245	0,3235	0,3225
0,74	0,3215	0,3204	0,3194	0,3184	0,3174	0,3164	0,3154	0,3144	0,3133	0,3123
0,75	0,3113	0,3103	0,3092	0,3082	0,3071	0,3061	0,3051	0,3040	0,3030	0,3019
0,76	0,3009	0,2999	0,2988	0,2978	0,2967	0,2956	0,2946	0,2935	0,2925	0,2914
0,77	0,2903	0,2893	0,2882	0,2871	0,2861	0,2850	0,2839	0,2828	0,2818	0,2807
0,78	0,2796	0,2785	0,2774	0,2763	0,2753	0,2741	0,2731	0,2720	0,2709	0,2698
0,79	0,2687	0,2676	0,2664	0,2653	0,2642	0,2631	0,2620	0,2609	0,2598	0,2587
0,80	0,2575	0,2564	0,2553	0,2542	0,2531	0,2519	0,2508	0,2497	0,2485	0,2474
0,81	0,2462	0,2451	0,2440	0,2428	0,2417	0,2405	0,2394	0,2382	0,2371	0,2359
0,82	0,2348	0,2336	0,2324	0,2313	0,2301	0,2292	0,2278	0,2266	0,2255	0,2243
0,83	0,2231	0,2220	0,2208	0,2196	0,2184	0,2172	0,2160	0,2149	0,2137	0,2125
0,84	0,2113	0,2101	0,2089	0,2077	0,2065	0,2053	0,2041	0,2029	0,2017	0,2005
0,85	0,1993	0,1981	0,1969	0,1957	0,1944	0,1932	0,1920	0,1908	0,1896	0,1884
0,86	0,1871	0,1859	0,1847	0,1834	0,1822	0,1810	0,1797	0,1785	0,1773	0,1760
0,87	0,1748	0,1735	0,1723	0,1711	0,1698	0,1686	0,1673	0,1661	0,1648	0,1635
0,88	0,1623	0,1610	0,1598	0,1585	0,1572	0,1560	0,1547	0,1534	0,1522	0,1509
0,89	0,1496	0,1484	0,1471	0,1458	0,1445	0,1432	0,1419	0,1407	0,1394	0,1381
0,90	0,1368	0,1355	0,1342	0,1329	0,1316	0,1303	0,1290	0,1277	0,1264	0,1251
0,91	0,1238	0,1225	0,1212	0,1199	0,1186	0,1173	0,1159	0,1146	0,1133	0,1120
0,92	0,1107	0,1094	0,1080	0,1067	0,1054	0,1040	0,1027	0,1014	0,1000	0,0987
0,93	0,0974	0,0960	0,0947	0,0933	0,0920	0,0907	0,0893	0,0880	0,0866	0,0853
0,94	0,0839	0,0826	0,0812	0,0798	0,0785	0,0771	0,0758	0,0744	0,0730	0,1717
0,95	0,0703	0,0689	0,0676	0,0662	0,0648	0,0634	0,0621	0,0607	0,0593	0,0579
0,96	0,0565	0,0552	0,0538	0,0524	0,0510	0,0496	0,0482	0,0468	0,0454	0,0440
0,97	0,0426	0,0412	0,0398	0,0384	0,0370	0,0356	0,0342	0,0328	0,0314	0,0300
0,98	0,0286	0,0271	0,0257	0,0243	0,0230	0,0214	0,0201	0,0186	0,0172	0,0158
0,99	0,0140	0,0129	0,0115	0,0101	0,0086	0,0072	0,0058	0,0043	0,0029	0,0014



2. Основные принципы официальной статистики в регионе Европейской экономической комиссии

Приняты в ходе 47-й сессии Европейской экономической комиссии ООН 15 апреля 1992 года во Дворце Наций в Женеве

Европейская экономическая комиссия,

принимая во внимание, что официальная статистическая информация является необходимой основой для развития в экономической, демографической, социальной и экологической областях, а также для взаимного познания и торговли между государствами и народами региона,

учитывая, что степень доверия общественности к официальной статистической информации в значительной мере зависит от уважения основополагающих ценностей и принципов, лежащих в основе любого демократического общества, стремящегося к самопознанию и уважению прав своих членов,

принимая во внимание, что качество официальной статистики и тем самым качество информации, предоставляемой правительству, экономике и общественности, в значительной мере зависит от сотрудничества граждан, предприятий и других респондентов в предоставлении надлежащих данных, требующихся для подготовки необходимой статистической информации,

ссылаясь на общие положения и нормы, принятые с этой целью в Европейской конвенции о защите прав человека, в Конвенции Совета Европы о защите прав отдельных лиц в связи с автоматизированной обработкой персональных данных от 28 января 1981 года, в Заключительном акте Хельсинского Совещания по безопасности и сотрудничеству в Европе и в Парижской хартии для новой Европы,

напоминая об усилиях правительственных и неправительственных организаций, занимающихся вопросами статистики, по разработке стандартов и концепций, позволяющих проводить сопоставления между странами,

ссылаясь также на Декларацию о профессиональной этике Международного статистического института,

учитывая консенсус, достигнутый в рамках Конференции европейских статистиков по вопросу о необходимости определения принципов, регулирующих деятельность государственных статистических учреждений в регионе и в государствах-членах,

принимает настоящую резолюцию:

I. Официальная статистика является необходимым элементом информационной системы демократического общества, обеспечивая правительство, экономику и общественность данными об экономическом, демографическом, социальном и экологическом положении. С этой целью социальные статистические данные, имеющие практическую ценность, подготавливаются и распространяются на объективной основе государственными статистическими учреждениями для обеспечения уважения права граждан на общественную информацию.

II. В целях сохранения доверия к официальной статистике статистические учреждения в соответствии со строго профессиональными соображениями, включая научные принципы и профессиональную этику, должны принимать решения в отношении методов и процедур сбора, обработки, хранения и представления статистических данных.

III. Для облегчения правильной интерпретации данных статистические учреждения должны предоставлять информацию в соответствии с научными стандартами в отношении источников, методов и процедур в области статистики.

IV. Статистические учреждения имеют право комментировать неправильную интерпретацию или неправильное использование статистических данных.

V. Данные для статистических целей могут собираться из всех типов источников, будь то статистические обследования или административная отчетность. Статистические учреждения должны выбирать источник с учетом качества, своевременности, затрат и бремени, которое ложится на респондентов.

VI. Персональные данные, собираемые статистическими учреждениями для подготовки статистической информации, независимо от того, относятся ли они к физическим или юридическим лицам, должны носить строго конфиденциальный характер и использоваться исключительно для статистических целей.

VII. Законы, нормы и меры, в рамках которых функционируют статистические системы, должны предаваться гласности.

VIII. Для обеспечения согласованности и эффективности в статистической системе необходимо осуществлять координацию деятельности статистических учреждений на уровне стран.

IX. Использование статистическими учреждениями в каждой стране международных понятий, классификаций методов способствует обеспечению согласованности и эффективности статистических систем на всех официальных уровнях.

X. Двустороннее и многостороннее сотрудничество в области статистики содействует улучшению систем официальной статистики во всех странах.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

А

Автокорреляция 492, 493
Анализ кластерный 193
  метод «ближайшего соседа» 203, 206, 207, 209, 211
  метод «дальнего соседа» 203, 206, 207, 209, 211
Асимметрия 161, 162, 163, 164, 170
  левосторонняя 162, 170
  правосторонняя 162, 170
Ассортимент 555, 558, 562

В

Вариация 120, 140, 141
  показатели 154
    размах, или амплитуда 154
    среднее квартильное расстояние 158
    среднее квадратическое отклонение 155, 156, 181
    среднее линейное (абсолютное) отклонение 155
    относительные показатели асимметрии 162
    предельно возможные значения 165, 166, 167
    эксцесс распределения 163
    моменты распределения 160
    центральные 160, 161
    плотность распределения 149, 150
Вариационный ряд 142
  варианта 142
  дискретный 142
  интервальный 143
  кумулятивный 147, 149
  ранжированный 142
  структурные характеристики
    децили 146
    квартили 151
    квинтили 152
    медиана 150
    мода 152
    перцентили 152
    частота 142
    частость 148
Выборка 214
  малая 250
  ошибки 223
    средняя ошибка выборочной средней 225
    — — доли 230
    предельная ошибка выборочной средней 229
    — — доли 230
    средняя ошибка разности двух выборочных средних 292, 316
  репрезентативность 213, 217
  способы оценки генеральных параметров 336, 358
  способ отбора 219
    бесповторный 219
    квотный 222
    механический 221
    многоступенчатый 219, 239, 244
    многофазовый 219, 244
    повторный 219
    серийный 219
    случайный 219, 220
    типический (стратифицированный, или районированный) 231

Г

Гало (слой) 26, 158, 164
Гипотеза 271
  альтернативная 273
  непараметрическая 271
  нулевая 271
  параметрическая 271
  простая 271
  сложная 271
проверка статистических гипотез 272, 474

этапы 272
гипотезы о законе распределения 274
о нормальном распределении 276, 277, 278, 279, 280
о распределении Пуассона 282, 283
о связи на основе критерия хи-квадрат 287, 288, 289, 290, 291, 292, 422
о средних величинах 292, 293, 294, 296, 297
коэффициента корреляции z-преобразования 348
коэффициента регрессии t-критерия 305, 306, 307, 308
Графики 106, 107
гистограмма 147
граф связей 569, 570
график сезонной волны 501, 502
дендрограмма 206
диаграммы 106
картограммы 106, 115, 116, 117, 119
картодиаграммы 106, 115, 118, 119
кривая Лоренца 615
кумулята 148, 149
огива 149
полигон 106, 147
радиальная диаграмма 501, 502
фигурные 112
Группировка 172
аналитическая 177, 182, 212
комбинационная 174
многомерная 176
многофакторная 186, 301
однофакторная 297
простая (монотетическая) 175
сложная (политетическая) 175
структурная 177, 180, 212
типологическая 177, 212

Д
Данные 22
анализ 28
временные 22
достоверность 49, 50
панельные 22
представление 28, 29
пространственные 22
сбор 27
сопоставимость 50
сравнимость 51
Дендрограмма 206
Диаграмма 105, 106, 108, 109, 111, 112, 504, 542, 543, 616
Динамика 108, 576
ряд динамики (временной) 445, 447
уровень ряда 445, 447
вид динамического ряда 445
интервальный 445, 447
моментный 445, 447
тенденция 448, 468
аналитическое выравнивание 468
многократное выравнивание 482, 483, 484, 513, 514
метод скользящей средней 503, 504
темп роста 87, 450, 452
цепной 450, 453
базисный 450, 452, 453
темп прироста 87, 454
цепной 454
базисный 454
абсолютное значение 1%-ного прироста 454
абсолютный прирост 450, 452
цепной 450
базисный 450
индекс Доу-Джонса 455, 456
средние показатели тенденции 459
средний уровень ряда 460, 461
хронологическая средняя 461
средний абсолютный прирост 461, 469
средний темп роста 463, 465, 467
средний темп прироста 463
Дисперсионный анализ 295, 407
однофакторный 296, 297, 298, 299
двухфакторный 299, 300, 301, 302, 303, 304

Дисперсия 156, 431
альтернативного признака 230
общая 184, 330, 331, 373
остаточная 184, 373
факторная 184
внутригрупповая 184, 330
межгрупповая 184, 330

Е
ЕГРПО 65, 71, 75
Единица наблюдения 56,79
— совокупности 19, 24

З
Закон 138
Закон больших чисел 138
Закономерность 16, 138, 139
динамическая 17
статистическая 17, 18

И
Измерители 20
натуральные 20
стоимостные 20
условно-натуральные 20
Индексы 526, 527, 528
агрегатные 537, 538, 539, 594
индивидуальные 528, 529
средние из индивидуальных 529, 531, 534, 536, 537
базисные 547
цепные 547
структуры 548, 549, 561
переменного состава 548, 550
постоянного состава 548, 550
свойства 545
включения-выключения 547
пропорциональности 547
соизмеримости 547
тест кружного испытания 546
тест обратимости во времени 545
тест обратимости по факторам 546
формы 537
Ласпейреса 533, 537, 590, 595
Пааше 536, 537, 595
Фишера 547
Интервал 144
группировочного признака 176
специализация интервалов 179
границы интервала 176,178
Информация 48, 428
Источник данных 49, 54
документальный 54
непосредственное наблюдение 54
опрос 54, 55

К
Кластер 193
Кластерный анализ 193
Классификация 11
алгомеративно-иерархическая 197, 211
многомерная 11, 190, 193
неиерархическая 211
Колебания 488, 510, 522
равные 175, 211
неравные 175, 211
типы 486, 487
долгопериодические ц
ские 485, 486, 487, 491
пилообразные (маятниковые) 480, 487, 492
сезонные 496, 497, 499, 501, 502
силы колебаний уровней ряда 489
среднее абсолютное отклонение 490, 524
среднее квадратичное отклонение 490, 507, 524
коэффициент колеблемости 489, 491, 507
Колеблемость случайная 486, 487, 488, 491
Контроль данных 71
логический 72, 73
счетный 72
Корреляция 323, 389, 521
гиперболическая 361
линейная 371, 376

нелинейная 358, 361
параболическая 358
парная 371, 376, 390
частная 378, 390
множественная 374, 381, 390
Корреляционное отношение 354
случайное 191
значение корреляционного отношения 190, 363
теоретическое 354
эмпирическое 184, 185, 189, 354
Коэффициент 516
абсолютного структурного сдвига 623, 624, 625
автокорреляции 488, 492, 493
вариации рангов 494, 495, 524, 626
множественный 372, 373, 374
парный (линейный) 371, 390, 423
частный 379, 390
Гатева 624
детерминации 158, 339, 340, 346, 372, 376, 421, 431, 628
Джини 615
изменения ранга долей 625, 626, 627
корреляции 516, 519
концентрации 167, 617, 619, 620
линейный коэффициент ранга долей 625
регрессии 336, 338, 519
осцилляции 158
эластичности 87, 366
Кумулята 106, 147, 148
Критерий 272, 422, 500, 620
F-критерий 293, 294, 296, 297
W-критерий Вилкоксона 308
D-критерий Колмогорова—Смирнова 284, 285, 286
Критическая дата учета 58
Критический момент наблюдения 58

М

Метод «ближайшего соседа» 204, 206, 207, 209, 210
Метод «группового соседа» 209, 211
Метод «дальнего соседа» 204, 205, 207, 209, 211
классификации 208, 211
Метод наименьших квадратов 336, 358, 362, 403, 476
двойной 405, 410
косвенный 401, 410
Метод многократного скользящего выравнивания 482
Меры
связи 370, 416, 434
сходства 209
различия 209
Многомерная структура 609
Модель сезонности 503, 504, 506, 508
Модели рекурсивные 190, 193, 393

Н

Наблюдение 49, 78, 215
выборочное 53, 215
моментное 254
единовременное 51
сплошное 52, 215
смешанное 215
статистическое 51
непрерывное (текущее) 51
периодическое выборочное 53
Нечеткое множество 164

О

Организации 31
международные 37, 39, 40, 43, 47
региональные (местные) 31, 36, 43
центральные 31, 33
Отчетность 63
унифицированная 66, 67
Ошибка 71
абсолютная допустимая 253, 254, 255
относительная 253
репрезентативности 223
систематическая 71, 72
случайная 72, 73
Оценка генеральных параметров
доли 229, 230
интервальная 229, 230
средней 227, 228
точечная 229, 230

П

Перепись 79
населения 58, 60
экономическая 77
Показатель 81, 83, 97
атрибуты 82, 97
система 82, 92
иерархической древовидной структуры 599, 600, 602
балансовой структуры 603
изменения структуры 621
абсолютные 84, 85, 621
относительные 84, 86, 90, 621, 624
ранговые 625, 626, 627
колеблемости 489, 491
концентрации 618, 619
монополизации 617
силы связи 182
специализации 617
тесноты связи 182
устойчивости 18, 489, 491
Поле корреляции 106, 328, 345
Правило разложения дисперсии (сложения дисперсий) 183, 330
Предприятие 53, 75, 76, 94, 140
Признак 22, 30, 83
классификация 22
альтернативный (дихотомический) 24, 213, 416
вторичный 23, 126
группировочный 173, 176, 211
опознавательный 60
специализация группировочных признаков 77, 177
дискретный 24
интервальный 25
количественный 23
косвенный 24
моментный 25
непрерывный 25
описательный 22
первичный 23, 124
экзогенный 393, 409, 569
эндогенный 392, 409, 569
Полигон распределения 106, 147
Признаковое пространство 191, 210
Прогнозирование 511, 512, 515
Программа наблюдения 59

Р

Расстояние 199, 200, 204, 205, 206
евклидово 194, 197, 198
Распределение
Пуассона 280, 281, 282
равночастотное (равномерное) 142, 144, 618
безусловное 429
условное 429
нормальное 155, 162, 163, 164, 276
асимметричное 187
стандартизированное 188, 189
Репрезентативность 213, 217
Ряды динамики 445
колебания 448, 489, 510, 522
корреляция 511, 516, 518, 520, 521
регрессия 328, 511, 571
линейная 335
нелинейная 358, 361
множественная 364, 367, 388
парная 335, 349, 361
по отклонениям от трендов 518, 520
по первым разностям 521
по уровням временных рядов 518, 524

С

Связь 320, 321
корреляционная 320, 564
статистическая (стохастическая) 320, 564

функциональная (жестко детерминированная) 320, 321, 564, 570, 597
Сила 183, 186, 338
теснота 338
Совокупность 21, 29, 121
выборочная 216
генеральная 216
гипотетическая 217
общая 26
реальная 217
частная 26, 29
Средняя величина 120, 121, 122, 169
квадратическая 134, 169
геометрическая 135, 169
кубическая 137, 169
гармоническая 136, 169
стандартизированная 189
степенная 137
арифметическая 123
математические свойства 128, 129, 130
многомерная 192, 193, 211
простая 124, 125, 131, 132, 133, 531
взвешенная 125, 131, 132, 133, 532
Среднее линейное отклонение 154, 155
Среднее квадратическое отклонение 154, 155
Средняя квадратическая ошибка выборочной средней 232, 324
— выборочной доли 224, 225, 232
— разности двух выборочных средних 616, 617
— разности двух выборочных долей (относительных величин) 229, 616, 617
Стандарты международные 42
Статистика 12, 13
гендерная 77
дескриптивная, *см. описательная статистика* 217
математическая 15
населения 15
общая теория 15
описательная, *см. дескриптивная статистика* 217
прикладная 15
региональная 36
экономическая 15
Структура 597, 599, 603, 609, 617
абсолютный показатель изменения структуры 621
средний квадратический показатель изменения структуры 623
среднее относительное изменение долей 624
Структура совокупности 86

Т

Таблица 101, 119
простая 101
групповая 101, 102, 119
комбинационная 101, 102, 119
типовая 103, 119
элементы 102, 104, 105
Теория вероятностей 15, 29
Теснота связи признаков 338, 412
методы измерения 412
коэффициент детерминации 329, 339, 346
коэффициент корреляции К. Пирсона 339
коэффициент корреляции рангов Ч. Спирмена 437, 438, 439, 494, 495
коэффициент корреляции рангов М. Кендэла 440
множественная регрессия 364, 367, 388
парная линейная корреляция 339, 340, 341, 343, 344, 352, 353
парная линейная регрессия 342, 349
коэффициенты взаимной сопряженности
Г. Крамера 423, 427, 429
К. Пирсона 423, 425, 428
А. Чупрова 426, 427, 428
коэффициент ассоциации 417, 418, 420
коэффициент контингенции 419, 420
теоретико-информационные меры связей 429, 433
коэффициент нормированной информации 431, 433
систематизированный коэффициент информации 433
λ-Гутмана 434, 435, 444
τ-Гудмена—Краскала 436, 444
коэффициент конкордации 441, 442, 443, 444
Тренд 476, 493, 511, 512
расчет параметров 476
форма тренда 523
линейный 469, 476, 513
параболический 469, 477
экспоненциальный 470, 478, 479
логарифмический 471
степенной 471
гиперболический 471, 472
логистический 472

У

Уровни измерения переменных 411, 412, 413, 415, 416
Устойчивость 18
направленности изменений 493, 494
тенденции 496
показатели устойчивости тенденции 496

Ф

Форма собственности 173, 174
организационно-правовая 173, 174
ответа 60, 61

Ц

Ценз 53
Цензовый принцип организации учета 53

Ч

Частная корреляция 378, 390
Частное корреляционное отношение 378, 390
Частота 142, 276, 283
клеточная 285
маргинальная 285
кумулятивная (накопленная) 149
Частость 148
Число групп 143, 174

Ш

Шкала 411
порядковая (ординальная) 412, 413, 414
номинальная 412
интервальная 411
отношений 411

Э

Экстраполяция 511
Эксперимент 56
Эксцесс 162, 163
Элементы таблицы 101
подлежащее 101
сказуемое 101
Эмпирическое корреляционное отношение 183, 184, 186, 187, 189, 354
Энтропия распределения 431

Я

Явление
массовое 17, 158, 163
Ядро 26, 29, 158, 164

Учебное издание

Елисеева Ирина Ильинична
Юзбашев Михаил Михайлович

ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ

Заведующая редакцией *Л. А. Табакова*
Редактор *Н. А. Кузнецова*
Младший редактор *Н. А. Федорова*
Художественный редактор *Ю. И. Артюхов*
Технический редактор *В. Ю. Фотиева*
Корректоры *Т. М. Колпакова, Н. Б. Вторушина, Г. В. Хлопцева*
Оформление художника *Н. М. Биксентеева*

ИБ № 4311

Сдано в набор 27.02.2003. Подписано в печать 09.10.2003.
Формат 60 × 88 1/16. Печать офсетная
Усл. п.л. 40,18. Уч.-изд. л. 35,15. Гарнитура «Таймс».
Тираж 5000 экз. Заказ № 3325. «С» 227

Издательство «Финансы и статистика»
101000, Москва, ул. Покровка, 7
Телефон: (095) 925-35-02,
факс (095) 925-09-57
E-mail: mail@finstat.ru http://www.finstat.ru

ГУП «Великолукская городская типография»
Комитета по средствам массовой информации
Псковской области,
182100, Великие Луки, ул. Полиграфистов, 78/12
Тел./факс: (811-53) 3-62-95
E-mail: VTL@MART.RU