Канцедал Сергей Андреевич

Доктор технических наук, профессор.
Специалист в области решения экстремальных задач дискретной математики.
Имеет многолетний опыт создания прикладного программного обеспечения для ЭВМ.
Автор учебников и учебных пособий для студентов высших и средних специальных учебных заведений, в том числе «Алгоритмизация и программирование», «Дискретная математика».

ISBN 978-5-8199-0439-8

Основы статистики

С.А. Канцедал

С.А. Канцедал

Профессиональное образование

Основы статистики

Издательский дом ФОРУМ

С. А. Канцедал

ОСНОВЫ СТАТИСТИКИ

Рекомендовано методическим советом Института искусств и информационных технологий в качестве учебного пособия для студентов средних специальных учебных заведений, обучающихся по группе специальностей «Экономика и управление»

Москва
ИД «ФОРУМ» — ИНФРА-М
2011

УДК 311(075.32)
ББК 65.051я723
К19

Рецензенты:
доктор технических наук, профессор кафедры математического моделирования Житомирского государственного технологического университета *А. В. Панишев*;
доктор технических наук, профессор, зав. кафедрой «Информатика и программное обеспечение вычислительных систем» Московского государственного института электронной техники (Технического университета) *Л. Г. Гагарина*

Канцедал С. А.

К19 Основы статистики: учебное пособие / С. А. Канцедал. — М.: ИД «ФОРУМ»: ИНФРА-М, 2011. — 192 с.: ил. — (Профессиональное образование).

ISBN 978-5-8199-0439-8 (ИД «ФОРУМ»)
ISBN 978-5-16-004362-3 (ИНФРА-М)

В книге на элементарном уровне изложены классические разделы описательной и аналитической статистики, а также проблемы принятия статистических решений в условиях риска и неопределенности. Все излагаемые задачи статистики сопровождаются многочисленными примерами, что существенно облегчает понимание студентами излагаемого теоретического материала. Особое внимание уделено описанию современных компьютерных технологий решения этих задач.

Книга предназначена в качестве учебника для учащихся экономических колледжей и не требует знаний, выходящих за пределы школьного курса математики. Она может быть использована также студентами вузов для ознакомления с предметом.

УДК 311(075.32)
ББК 65.051я723

ISBN 978-5-8199-0439-8 (ИД «ФОРУМ»)
ISBN 978-5-16-004362-3 (ИНФРА-М)

Подписано в печать 27.04.2010. Формат 60×90/16.
Печать офсетная. Гарнитура «Таймс». Усл. печ. л. 12,0. Уч.-изд. л. 12,5.
Бумага офсетная. Доп. тираж 500 экз. Заказ № 3325.

Отпечатано с готовых диапозитивов в ОАО ордена «Знак Почета» «Смоленская областная типография им. В. И. Смирнова».
214000, г. Смоленск, проспект им. Ю. Гагарина, 2.

Предисловие

С самых давних времен для изучения различных явлений природы и общества люди вели наблюдения, ставили эксперименты, проводили опросы и опыты. Результаты этих действий представлялись числовыми и качественными данными и рассматривались как случайные события и величины, которые затем интерпретировались тем или иным способом.

Постепенно в результате этой деятельности сформировалось научное направление, которое с течением времени трансформировалось в отдельную отрасль науки — статистику.

Пионерами статистики были европейские математики: У. Петти, И. Бернулли, Т. Байес, У. Госсет, К. Пирсон, Р. Фишер, С. Крамер и многие другие.

У. Петти, например, составил первый отчет об уровне смертности в Лондоне. И. Бернулли дал правило определения вероятности успешных исходов в серии независимых испытаний. К. Пирсон обосновал известное и широко используемое в статистике распределение вероятностей χ^2. С. Крамер оставил после себя фундаментальную книгу «Математическая статистика».

По современным воззрениям статистика рассматривает способы получения (сбора), обработки, анализа и истолкования данных о явлениях природы и общества, а также методы использования результатов анализа для принятия рациональных решений в той или иной человеческой деятельности.

Как наука статистика состоит из трех взаимосвязанных разделов: описательной статистики, аналитической статистики и теории принятия статистических решений.

Описательная статистика ограничивается способами получения полной или частичной (выборочной) информации о наблюдаемых явлениях, объектах и процессах, ее анализом, истолкованием, а также методами представления полученных данных в удобной табличной и графической формах.

Аналитическая статистика — это методы получения статистических заключений, характеризующих наблюдаемые явления,

процессы или объекты как бы на основании полного набора данных, располагая, однако, только выборочной информацией, полученной на стадии описательной статистики.

В отличие от описательной и аналитической статистики теория статистических решений представляет собой относительно новый раздел статистики, посвященный изучению методов принятия решений в ситуациях, которые в различной мере описываются статистическими данными.

С определенной степенью полноты материал, входящий в указанные разделы статистики, изложен в настоящем учебном пособии.

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Глава 1
РЯДЫ РАСПРЕДЕЛЕНИЯ ЧАСТОТ И ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВЫБОРКИ

1.1. Основные термины и определения

Каждая научная дисциплина использует свои специфические термины или, как принято говорить, свой язык. Не всем изучающим эту дисциплину они понятны, в связи с чем под рукой необходимо иметь энциклопедический словарь или использовать Интернет. Чтобы избавить читателя от занятий листать словарь и входить в Интернет, приведем основные термины и определим понятия, которые будут фигурировать в дальнейшем изложении.

Начнем с простейшего. **Случайное явление** — это такое явление, исход которого (практическое его проявление) заранее не может быть предсказан.

Когда мы вытягиваем одну карту из хорошо тасованной колоды карт, наперед нельзя сказать, какой она будет масти. Когда мы приходим на остановку нескольких трамваев и не знаем расписания их движения, нельзя угадать, какой трамвай подойдет. Когда судья перед капитанами футбольных команд подбрасывает и ловит монету, нельзя заранее сказать, выпадет орел или решка и, таким образом, нельзя предугадать, какие ворота достанутся той или иной команде. Когда вы просите продавца магазина взвесить 0,5 кг ветчины, нельзя заранее угадать, насколько он ошибется.

Все это примеры случайных явлений, которые, вообще говоря, легко умножить. Таким образом, отличительная черта случайности — непредсказуемость исхода априори, т. е. до опыта.

Явление называется **детерминированным**, если его исход заранее определен. Например, мы знаем, что за ночью всегда наступит утро, затем день, вечер и снова ночь.

Как уже говорилось, для изучения явлений природы и общества люди проводят опыты, опросы и наблюдения.

Испытание или опыт — это комплекс процедур, которые можно повторять сколько угодно раз при одних и тех же условиях без изменения. Результат испытаний, т. е. всякий факт, который может произойти или не произойти, называется **случайным событием**. **Наблюдение** представляет собой фиксацию исхода события или явления.

Например, в опыте подбрасывания монеты, повторяемом двадцать раз подряд, может произойти или не произойти событие — выпадение орла. Наблюдениями исходов в этом случае будут фиксации того факта, что же выпало — орел или решка.

Те события, которые в результате каждого испытания происходят неизбежно, называются **достоверными**. Например, сколько бы раз мы ни бросали камень весом 0,3 кг вверх, он обязательно упадет на землю. События, которые в результате всего множества опытов не происходят никогда, называют **невозможными**. Например, выпадение более шести очков при бросании игральной кости — событие невозможное.

События **равновозможные** (одинаково возможные), если есть основание полагать, что появление одного из них в одном и том же опыте не более возможно, чем появление другого. Равновозможно появление орла или решки при бросании монеты, выпадении трех или шести очков при бросании игральной кости и т. д.

События называются **несовместными**, если появление одного из них в одном и том же опыте исключает появление другого. Так, появление герба при бросании монеты исключает появление решки. Такие события часто называют **взаимоисключающими**.

Предположим, осуществляются наблюдения некоторого явления или исходов опыта. Например, ежедневная фиксация «дождь или сухо» в течение года, «орел или решка» при 100 бросаниях монеты, количество очков от одного до шести при 30 бросаниях игральной кости. В течение этих наблюдений фиксируется, какое же из возможных событий и сколько раз произошло.

Число произошедших событий называется его **частотой**. Отношение числа появлений n_A конкретного события A к общему числу наблюдений n, т. е. $W_n(A) = n_A/n$, называется *относительной частотой* (частостью) случайного события A.

Так, если при 50 подбрасываниях монеты событие A — выпадение орла — наблюдалось 22 раза, частота равна 22, а частость этого события $W_{50} = 22/50 = 11/25 = 0{,}44$.

Относительная частота достоверного события всегда равна единице, так как происходит это событие в каждом опыте, вследствие чего $n_A = n$. Относительная частота невозможного события равна нулю, так как это событие не происходит никогда, т. е. $n_A = 0$. Поэтому частость случайного события A лежит в пределах $0 \le W_n(A) \le 1$.

Статистические данные, которые характеризуют явления и получают в результате опытов, представляют собой **случайные величины**, т. е. величины, которые могут принимать то или иное значение — заранее неизвестно какое.

Различают величины двух типов: *дискретные* (прерывистые) и *непрерывные*. По существу это переменные, которые получают значения в результате счета или измерения. В геометрической интерпретации на числовой оси дискретные переменные — это целые числа, отображаемые отдельными точками, непрерывные величины — интервалы, в которые попадают в общем случае рациональные числа, полученные в результате измерения. Непрерывные переменные всегда ограничены определенными пределами.

Предположим, по некоторому предмету ученик сдает экзамен. Оценки, которые он может получить, — 2, 3, 4, 5 — дискретная случайная величина, принимающая целые значения 2, 3, 4, 5. Число избирателей, которые могут прийти на данный избирательный участок, — дискретная случайная величина, ограниченная слева нулем, справа количеством избирателей, внесенных в список. Число очков 0, 1, 2, которые может получить данная футбольная команда в результате очередной игры, — дискретная случайная величина.

Время проезда на автомобиле из города A в город B — непрерывная случайная величина, принадлежащая интервалу $[t_{min}, t_{max}]$. Погрешность Δ приближенного алгоритма, предназначенного для решения некоторой задачи, — непрерывная случайная величина, лежащая в интервале $[0, \Delta_{max}]$.

Месячный доход семей некоторой категории служащих данного региона — непрерывная случайная величина, находящаяся в интервале $[d_{min}, d_{max}]$.

Приведенные примеры случайных величин, безусловно, легко умножить.

Изучая случайные явления и проводя опыты, принципиально можно провести все возможные наблюдения явлений и все опыты. В этом случае говорят о **сплошном** изучении и о том, что рассматривается **генеральная совокупность** наблюдений.

Когда же имеют дело с частью наблюдений явлений и результатов опытов, говорят о **выборочном** изучении и о **выборке** наблюдений.

Например, руководство некоторого университета интересуют данные о росте ребят первого курса, образованного тремя группами студентов, численностью 75 человек. Можно измерить рост каждого студента, отметить минимальный, максимальный и вычислить средний рост, а также получить другие числовые характеристики. Это будет сплошным изучением роста, а генеральная совокупность — 75 студентов. Выборочное изучение — такое изучение, когда для измерения роста будет отобрана часть студентов, например 25. В этом случае именно число 25 и составляет **объем** выборки.

На основании измерения роста этого числа студентов получают характеристики: минимальный, максимальный, средний рост и др. Полученные на основании выборки данные называют **статистиками**. В рассматриваемом случае — это статистики роста студентов.

Статистики используют для оценки характеристик генеральной совокупности, которые принято называть **параметрами**. В этом и состоит одна их основных задач аналитической статистики: по характеристикам выборки оценить параметры генеральной совокупности. Иными словами, на основании части данных определить общие свойства изучаемого признака.

Такой подход практически обусловлен рядом обстоятельств. Не всегда возможно измерить характеристики каждого элемента генеральной совокупности, например в том случае, когда их число бесконечно или требует больших затрат времени, или измерение этих характеристик связано с большими финансовыми расходами, или когда в процессе измерения происходит разрушение или изменение характеристик элементов. Например, в случае контроля качества ламп накаливания, когда при проверке многие лампы выходят из строя.

Безусловно, в тех случаях, когда допустимо измерение характеристик каждого элемента генеральной совокупности, проводят сплошное изучение и ограничиваются статистическим описанием, опуская решение проблем статистического заключения. При выборочном изучении такое заключение необходимо проводить в обязательном порядке, так как нужно знать, насколько правдоподобны выводы относительно свойств генеральной совокупности, полученные на основании выборки.

1.2. Построение рядов распределения частот

Данные являются основой статистических исследований, ее фундаментом. По определению данные — это значения, которые присвоены конкретному наблюдению или измерению [1]. Достоверность данных определяет правдивость и объективность выводов, полученных на основании статистической обработки этих данных.

Данные принято классифицировать по разным направлениям. Так, по способу источников получения различают первичные и вторичные данные.

Первичные данные формирует лицо, непосредственно их использующее. Для этого проводят наблюдения, опросы, ставят эксперименты, фиксируют результаты измерений. Достоверность таких данных обеспечивается тем исследователем, который их собирает.

Вторичными данными являются те данные, которые собраны людьми, не проводящими дальнейших статистических исследований на основе этих данных. Главным недостатком этих данных является то, что способ их сбора не может быть проконтролирован. Поэтому во многих случаях достоверность таких данных может быть сомнительной.

Различают также количественные и качественные данные. Количественные данные позволяют проводить числовой анализ, в связи с чем они являются основой статистических методов исследования.

Качественные данные используют описательные выражения для рассматриваемых объектов. Например, имя опрашиваемого респондента, его возраст, семейное положение, пол и т. д.

Ряды распределения частот строят для облегчения анализа и толкования данных. Ряд распределения представляет собой перечень нескольких групп данных, для каждой из которых указано количество единиц изучаемого признака, — частота и частость $W_n(A)$.

В табл. 1.1 приведена выборка результатов измерения роста 50 студентов некоторого учебного заведения с точностью до 1 см.

Таблица 1.1. **Рост 50 студентов**

167	173	171	174	160	168	170	170	166	173
168	167	172	169	171	169	176	170	172	171
177	169	171	174	166	168	170	172	172	170
171	170	164	165	170	169	172	175	166	167
170	172	167	170	171	168	166	173	169	167

Для статистика, изучающего рассматриваемый признак студента, т. е. его рост, эта таблица малоинформативна. Его, как правило, интересуют некоторые общие черты случайности. Это прежде всего минимальный и максимальный рост студентов и соответствующий интервал между этими величинами. Как по этому интервалу рассредоточен рост студентов, равномерно или есть какие-то сгущения. Какова величина среднего роста как представителя данной выборки, а также другие характеристики. На все эти вопросы дает ответ ряд распределения.

Построение этого ряда начинают с вычисления диапазона колебаний — размаха вариаций R, в данном случае роста студентов. Для этого в исходной таблице данных находят минимальный и максимальный элементы x_{min}, x_{max} и вычитают из второго первое, т. е. получают $R = x_{max} - x_{min}$.

Далее решают вопрос о числе интервалов равной длины, на которые будет разбит диапазон R.

К сожалению, четких правил, которых следует придерживаться при решении этого вопроса, нет. Обычно считают, что правильно составленный ряд распределения должен содержать от 6 до 15 интервалов.

Предварительно число интервалов может быть определено по формуле Штюргеса $k = 1 + 3{,}32 \lg n$. После этого с учетом полученной величины диапазона R число интервалов уточняется.

В качестве примера определим количество интервалов для данных табл. 1.1.

Согласно приведенной формуле $k = 1 + 3{,}32 \lg 50 = 1 + 3{,}32 \times \times 1{,}699 = 6{,}641 \approx 7$, а $R = 177 - 160 = 17$. Исходя из этого примем количество интервалов равным 6, с длиной каждого из них 3 см. Тогда $6 \cdot 3 = 18 > 17$ покрывает размах вариации R, т. е. при группировке ни одно данное не будет потеряно.

Теперь, пользуясь длиной интервала 3 см, наименьшим значением $x_{min} = 160$ и количеством интервалов 6, построим их последовательность 160—162, 163—165, 166—168, 169—171, 172—174, 175—177, покрывающую размах вариации $R = 17$.

Заключительный этап построения ряда распределения состоит в подсчете количеств наблюдений, попадающих в каждый интервал, т. е. в вычислении частот и частостей. Результаты этих действий для рассматриваемого примера представлены в табл. 1.2.

Таблица 1.2. **Ряд распределения роста студентов**

№ группы	Рост студентов	Число студентов (частота)	Относительная частота (частость) $W_{50}(A)$
1	160—162	1	0,02
2	163—165	2	0,04
3	166—168	13	0,26
4	169—171	20	0,4
5	172—174	11	0,22
6	175—177	3	0,06

Из таблицы наглядно следует, каков минимальный и максимальный рост студентов и то, что их рост неравномерно распределен по диапазону R. Что преимущественный рост студентов, составляющий 40 % общего их числа, лежит в пределах 169—171 см, что студентов роста 166—168 см примерно такое же количество, как и студентов роста 172—174 см.

Таким образом, разобранный пример показывает, насколько информативнее становятся наблюдения после представления их в виде ряда распределения.

В рядах распределения выделяют нижние и верхние пределы групп данных, нижние и верхние их границы, а также средние точки, часто называемые метками. Эти величины используются при вычислении статистик выборки или вычислении параметров, если данные представляют генеральную совокупность наблюдений.

Нижние и верхние пределы определяют начало и конец каждой группы. Например, согласно табл. 1.2 нижний и верхний пределы третьей группы данных соответственно 166 и 168 см.

Границы групп — это точки, отделяющие одну группу от другой. Как правило, они расположены посередине между верхним пределом низшей группы и нижним пределом верхней группы данных. Таким образом, граница, разделяющая группы 169—171 и 172—174, расположена точно посредине между 171 и 172 см и, следовательно, равна 171,5 см.

Длина интервала определяется в начальной стадии построения ряда распределения. Когда же этот ряд построен, она может быть вычислена как разность между нижними и верхними пределами двух соседних групп либо как разность между нижней и верхней границей группы.

Например, длина интервала по пределам третьей и четвертой групп $i = 169 - 166 = 3$, по границам третьей группы $i = 168{,}5 - 165{,}5 = 3$. Средняя точка (метка) делит интервал пополам.

Следовательно, для третьей группы она равна $165{,}5 + i/2 = 165{,}5 + 1{,}5 = 167$.

Изложенное имело отношение только к непрерывным случайным величинам. Вместе с тем ряды распределения могут быть построены и для дискретных величин. Практически правила их конструирования остаются неизменными.

В качестве примера рассмотрим построение ряда распределения прогулов 16 работников некоторого предприятия, исходные данные которых представлены в табл. 1.3.

Таблица 1.3. **Прогулы работников в днях**

0	2	4	7	1	2	6	11	5	6	4	8	2	11	1	15

Согласно данным таблицы размах колебаний $R = x_{max} - x_{min} = 15 - 0 = 15$, количество интервалов по Штюргесу $k = 1 + 3{,}32 \lg 16 = 4{,}998 \approx 5$.

Примем это количество в качестве числа групп. Тогда в каждую группу попадет три наблюдения. Ряд распределения, построенный на этом основании, приведен в табл. 1.4.

Таблица 1.4. **Ряд распределения прогулов**

№ группы	Количество прогулов	Число работников	Относительная частота $W_{16}(A)$
1	0—2	6	0,375
2	3—5	3	0,1875
3	6—8	4	0,25
4	9—11	2	0,125
5	12—15	1	0,0625

Ряд распределения для дискретных величин отличается тем, что в нем из-за дискретности исходных данных не могут быть определены границы групп, а также их метки. Однако при практических расчетах условно можно считать, что они существуют.

1.3. Графические представления рядов распределения

Графические представления рядов распределения используют для того, чтобы повысить наглядность представления построенного ряда. Одним из простейших способов графического изображения ряда является **гистограмма**.

Гистограмма строится следующим образом. По горизонтальной оси в определенном масштабе откладываются отрезки, соответствующие длинам интервалов, обозначенных границами групп. Над каждым отрезком изображается прямоугольник с высотой, равной числу результатов наблюдений, попавших в соответствующий интервал, т. е. частота.

На рис. 1.1 изображена гистограмма для ряда распределения роста студентов, представленного табл. 1.2.

Весьма наглядным способом изображения распределения частот является полигон (рис. 1.2). Для его построения по горизонтальной оси откладывают метки групп, по вертикальной —

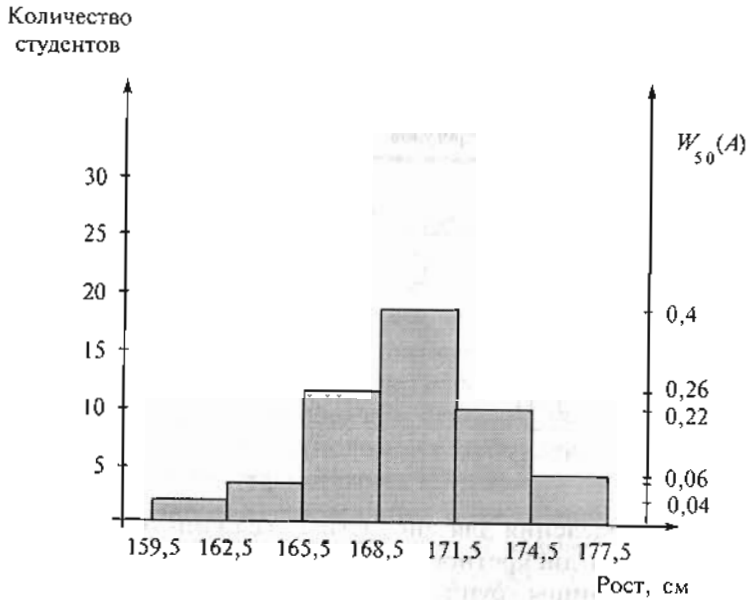

Рис. 1.1. Гистограмма роста студентов

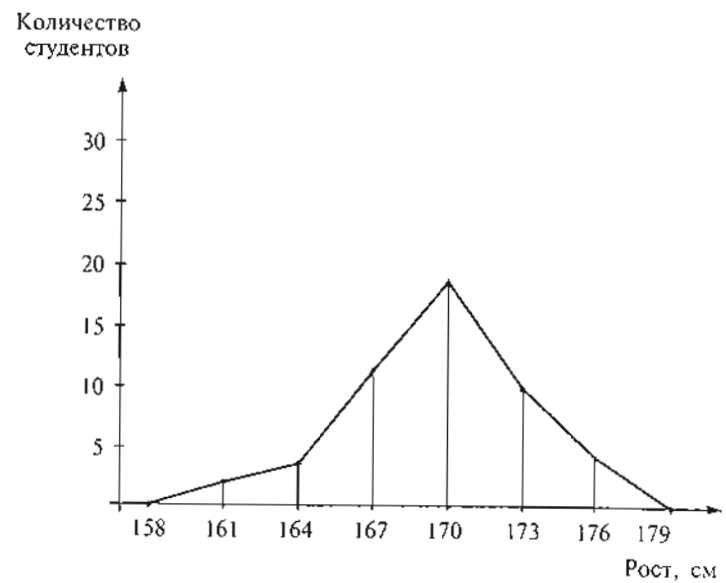

Рис. 1.2. Полигон роста студентов

частоты, соответствующие меткам. Затем полученные ординаты соединяют ломаной кривой.

На рис. 1.3 показано соотношение между гистограммой и полигоном.

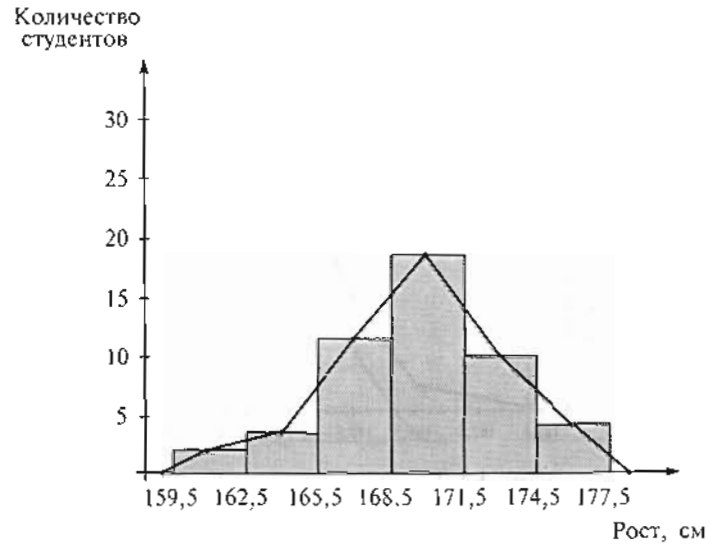

Рис. 1.3. Гистограмма и полигон роста студентов

Предполагая, что распределение наблюдаемого признака (в данном случае роста студентов) внутри каждого интервала равномерно и что полигон сглажен плавной кривой, считают, что он представляет собой частоту распределения этого признака внутри изучаемого диапазона.

Не менее важным способом графического представления данных является кривая накопленных (кумулятивных) частот. Кумулятивную частоту каждой очередной группы получают путем суммирования частоты этой группы с суммой частот предшествующих групп.

Таким образом, в рассматриваемом примере кумулятивная частота первой группы — 1, второй — 3, третьей — 16, четвертой — 36, пятой — 47, шестой — 50.

При графическом изображении кривой накопленных частот по горизонтальной оси откладывают границы интервалов, по вертикальной — соответствующие кумулятивные частоты. Далее

полученные точки вертикалей соединяют ломаной кривой и сглаживают. На рис. 1.4 изображена кривая накопленных частот роста студентов (кумулята).

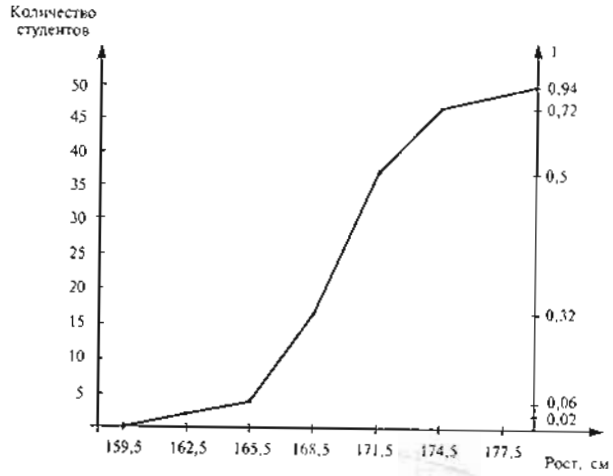

Рис. 1.4. Кумулята роста студентов

Имея в распоряжении кривую накопленных частот, легко подсчитать число наблюдений, лежащих в любом интервале изучаемого диапазона. Это число находится как разность ординат кривой в соответствующих точках. Например, число наблюдений интервала 165,5—174,5 равно 47 – 3 = 44.

С помощью кумулятивной кривой можно просто определить относительное число наблюдений, не превосходящих заданного числа или превосходящих это число. Для этого кумуляту необходимо представить в шкале относительных частот $W_n(A)$. Такая шкала показана в правой половине рис. 1.4.

Проводя горизонтальную линию, соответствующую некоторой относительной частоте, например 32 %, можно видеть, что это число обеспечивается не более чем 16 наблюдениями. Особый интерес представляет точка кривой, для которой 50 % частот наблюдений лежит слева и 50 % — справа. Эта точка называется *медианой*.

Форма кривой распределения частоты — сглаженный полигон — характеризуется двумя показателями: асимметрией и эксцессом. С помощью показателя асимметрии характеризуют степень ее отклонения от симметричной формы, при помощи показателя эксцесса измеряют ее островершинность.

Если левая половина кривой распределения частот не является зеркальным отображением правой ее половины, такая кривая является асимметричной и, следовательно, она отображает асимметричное распределение частот. Если же это не так — распределение считается симметричным.

Различают отрицательную и положительную асимметрии. На рис. 1.5 представлены асимметричные и симметричная кривые распределения.

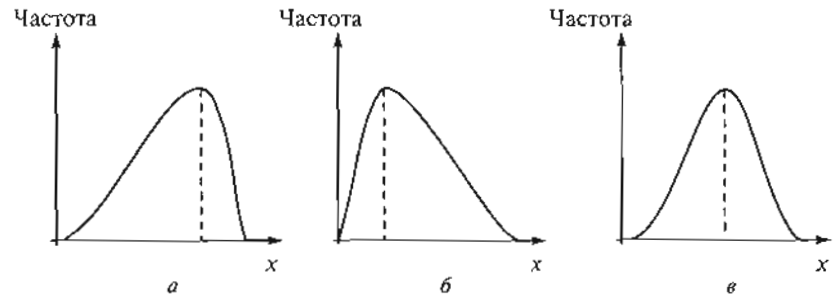

Рис. 1.5. Кривые распределения частот:
а — отрицательная асимметрия; *б* — положительная асимметрия; *в* — симметричная кривая

Концы кривой распределения принято называть «хвостами». Таким образом, если «хвост» асимметричного распределения лежит в области малых значений наблюдаемого признака — это отрицательная асимметрия. Если же «хвост» распределения лежит в области больших значений этого признака, налицо положительная асимметрия.

При оценке кривой распределения частот по характеристике эксцесс выделяют кривые с эксцессом меньше нормального, с нормальным эксцессом и с эксцессом больше нормального.

Кривые с плоской вершиной имеют эксцесс меньше нормального, с острой вершиной — больше нормального. Промежуточный уровень вершины определяет кривую с нормальным эксцессом. Перечисленные виды кривых изображены на рис. 1.6.

Кривые асимметрии показывают, в какую сторону смещается концентрация наблюдений. Кривые эксцесса определяют плотность рассеяния наблюдений. Если эксцесс меньше нормального,

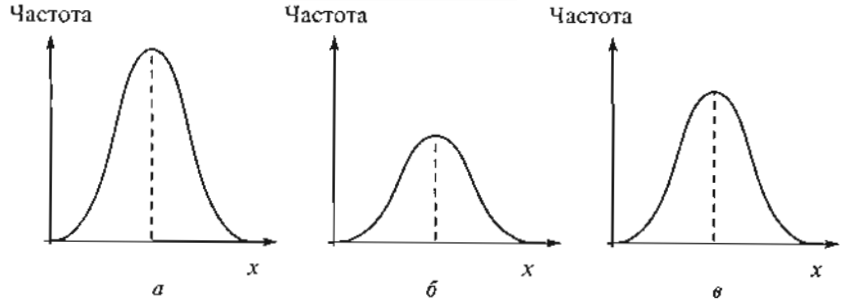

Рис. 1.6. Кривые распределения частот:
а — эксцесс больше нормального; *б* — эксцесс меньше нормального; *в* — нормальный эксцесс

наблюдения широко рассеяны по всему диапазону измерений, если больше нормального — они очень концентрированы возле определенной точки.

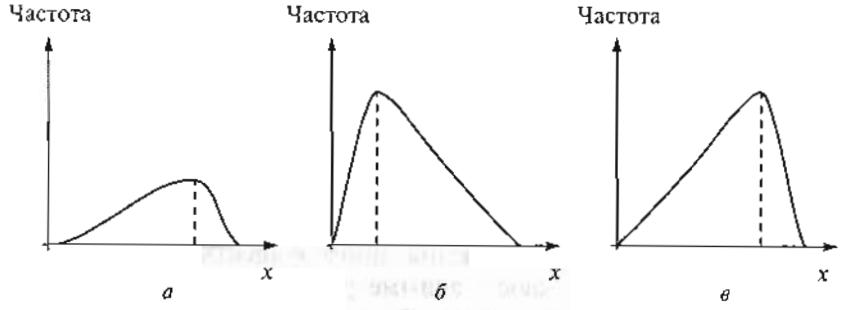

Рис. 1.7. Кривые распределения частот:
а — отрицательная асимметрия и эксцесс ниже нормального; *б* — положительная асимметрия и эксцесс выше нормального; *в* — отрицательная асимметрия и эксцесс больше нормального

Практически асимметрия и эксцесс кривых распределения изменяются одновременно. Такие кривые показаны на рис. 1.7.

1.4. Числовые характеристики центра распределения

Числовые характеристики выборки или генеральной совокупности в отличие от исходной таблицы данных, рядов распределения и наглядных их графических изображений представляют собой общие характеристики наблюдений, выраженные в числах.

К таким характеристикам относятся характеристики центра распределения или, как их часто называют, характеристики положения, и характеристики вариации (разброса) данных, как по всему изучаемому диапазону, так и в отношении центра распределения, а также характеристики асимметрии и закона распределения.

Центр распределения — это среднее значение, которое обобщает и представляет весь диапазон данных. В обыденной жизни мы часто пользуемся понятием среднего. Нередко говорим о средней скорости движения поезда, например, Москва—Кисловодск, о средней цене на хлеб, молоко, мясо в некотором регионе, о средней заработной плате работающих металлургов, шахтеров и т. д. При этом понимаем, что на некоторых перегонах поезд идет быстрее, чем на других, различные сорта хлеба, молока и мяса имеют разную цену, а работающие различных разрядов получают разную зарплату. Но путем применения среднего мы исключаем такие нюансы.

Аналогично понятие среднего используется и в статистике: частоты по всему диапазону различные, а их среднее характеризуют весь набор данных.

Среднее значение всегда расположено внутри диапазона данных. Поэтому в качестве числовых характеристик центра распределения используются различные средние: арифметическое, медиана, мода.

Среднее арифметическое может быть вычислено и по исходной таблице наблюдений, и по ряду распределения. Пусть x_1, x_2, ..., x_i, ..., x_n — значения n наблюдений, часто говорят вариант, представленные таблицей. Тогда среднее арифметическое $\overline{X}$ (среднее) определяется по формуле

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_i + \ldots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (1.1)$$

где знак $\sum_{i=1}^{n}$ означает суммирование данных от первого до n-го.

Иными словами, среднее арифметическое представляет сумму значений всех наблюдений (сумму вариант), деленную на их число. Для того чтобы получить эту сумму на основании табл. 1.1, не-

обходимо просуммировать ее элементы и разделить полученное значение на это число элементов.

При вычислении среднего арифметического на основании ряда распределения мы не располагаем элементами исходной таблицы. Есть только последовательность групп данных, их интервалов, число k этих интервалов, их средние точки (метки), а также частота признака для каждого интервала.

Пусть x_{1s}, x_{2s}, ..., x_{ks}, и f_1, f_2, ..., f_k, соответственно значения меток и частота меток интервалов. Тогда произведение $x_{1s} \cdot f_1$ — сумма значений наблюдений первого интервала, $x_{2s} \cdot f_2$ — сумма значений наблюдений второго интервала и т. д. На этом основании общая сумма значений наблюдений по всем интервалам равна

$$x_{1s}f_1 + x_{2s}f_2 + \ldots + x_{ls}f_l + \ldots + x_{ks}f_k = \sum_{l=1}^{k} x_{ls}f_l.$$

Для того чтобы получить среднее арифметическое всех значений вариант, необходимо эту сумму разделить на их число n. Поэтому формула для вычисления среднего арифметического на основании ряда распределения имеет такой вид:

$$\overline{X} = \frac{\sum_{l=1}^{k} x_{ls}f_l}{n}. \qquad (1.2)$$

Применим эту формулу для вычисления среднего арифметического ряда распределения, представленного в табл. 1.2.

Средние точки интервалов 1, 2, ..., 6 соответственно равны 161, 164, 167, 170, 173, 176. Частоты, взятые из табл. 1.2, имеют такие значения: 12, 13, 20, 11, ... Поэтому

$$\overline{X} = \frac{161 \cdot 1 + 164 \cdot 2 + 167 \cdot 13 + 170 \cdot 20 + 173 \cdot 11 + 176 \cdot 3}{50} \approx 170.$$

Необходимо отметить, что вычисление среднего арифметического на основании ряда распределения вносит некоторую погрешность в расчеты, так как значения наблюдений каждого интервала заменяются их средним значением. И чем интервал больше, тем больше погрешность.

Медиана представляет собой значение наблюдения, которое накопленную частоту делит пополам. Для того чтобы определить медиану для дискретных данных, необходимо упорядочить эти данные по возрастанию и найти среднюю варианту. Если число вариант n нечетное, то индекс медианы находится так:

$$I_m = \frac{n}{2} + \frac{1}{2}.$$

Например, пусть $n = 21$. Тогда

$$I_m = \frac{21}{2} + \frac{1}{2} = 10{,}5 + 0{,}5 = 11$$

и медиана определяется вариантой x_{11}. Если же число вариант четное, то медиана условно находится посередине между $\frac{n}{2}$ и $\frac{n}{2} + 1$ индексами вариант. Например, для $n = 20$ медиана находится между 10 и 11 вариантами. Аналогично можно действовать и для непрерывных данных.

При сгруппированных данных, представленных в виде ряда распределения, предварительно находят группу, содержащую медиану, часто говорят медианную группу, после чего непосредственно в этой группе устанавливают медиану.

Поскольку медиана делит накопленную суммарную частоту распределения пополам, медианная группа — это та группа данных, для которой накопленная частота x_n равна половине суммарной частоты или больше нее. Например, если суммарная частота распределения равна 50, то группа, для которой $f = 25$ или непосредственно следующая за ней — медианная.

Медиану в группе определяют методом интерполяции. Интерполяционная формула такая:

$$X_m = a_m + \frac{\frac{n}{2} - f_{m-1}^{k}}{f_m} d_m, \qquad (1.3)$$

где a_m — нижняя граница медианной группы;

n — суммарная кумулятивная частота;

f^k_{m-1} — кумулятивная частота группы данных, расположенной непосредственно перед медианной группой;

f_m — частота медианной группы;

d_m — интервал медианной группы.

В качестве примера вычислим медиану для ряда распределения, представленного в табл. 1.2. Для этого таблицу дополним данными о границах групп и накопленных частотах. В результате получим табл. 1.5.

Таблица 1.5. **Ряд распределения для расчета медианы**

№ группы	Рост студентов (пределы групп)	Рост студентов (границы групп)	Число студентов (частота)	Накопленная частота	Относительная частота $W_{50}(A)$
1	160—162	159,5—162,5	1	1	0,02
2	163—165	162,5—165,5	2	3	0,04
3	166—168	165,5—168,5	13	16	0,26
4	169—171	168,5—171,5	20	36	0,4
5	172—174	171,5—174,5	11	47	0,06
6	175—177	174,5—177,5	3	50	0,06

Определяем медианную группу.

Половина суммарной накопленной частоты $\frac{n}{2} = \frac{50}{2} = 25$. Группа, имеющая такую кумулятивную частоту, отсутствует. Группа с большей частотой — четвертая, т. е. $m = 4$. Далее вычисляем медиану.

Нижняя граница четвертой группы $a_4 = 168{,}5$. Кумулятивная частота предшествующей группы $f_3^k = 16$. Частота медианной группы $f_4 = 20$. Длина интервала $d_4 = 3$. Таким образом:

$$X_m = 168{,}5 + \frac{25 - 16}{20} \cdot 3 = 168{,}5 + 1{,}35 = 169{,}85.$$

Мода является третьей числовой характеристикой центра распределения. Она представляет собой наблюдение (варианту),

которая с наибольшей частотой встречается в ряду распределения. Если таких вариант больше одной, распределение называется бимодальным. В противном случае оно одномодальное. Если ни одна из вариант не повторяется, мода отсутствует.

Например, последовательность наблюдений 3, 5, 7, 10, 11, 13 моды не имеет. Наоборот, последовательность 3, 5, 6, 6, 6, 10, 13, 15 одномодальна и мода равна 6. Последовательность 3, 5, 5, 5, 6, 7, 7, 7, 10, 13 — бимодальна. Мода равна 5 и 7.

Моду можно найти по исходной таблице наблюдений и по ряду распределения. В первом случае поиск сводится к просмотру данных и отметке варианты, которая встречается в таблице чаще других вариант.

Во втором случае поиск моды начинается с выделения интервала с наибольшей частотой, так называемого модального интервала. Затем в его пределах методом интерполяции окончательно устанавливается мода. Интерполяционная формула может быть такой:

$$X_{md} = a_m + \frac{(f_m - f_{m-1})d_m}{(f_m - f_{m-1}) + (f_m - f_{m+1})}, \qquad (1.4)$$

где a_m — нижняя граница модального интервала;

f_{m-1}, f_m, f_{m+1} — частоты интервалов, предшествующие модальному, модального и следующего за модальным интервалом;

d_m — длина модального интервала.

Рассмотрим поиск моды на примерах данные табл. 1.1 и 1.5.

Просматривая данные табл. 1.1, находим, что мода $X_{md} = 170$. Это значение повторяется 8 раз.

Согласно данным табл. 1.5 модальный интервал четвертый, т. е. $m = 4$. Подставляя данные из табл. 1.5 в интерполяционную формулу (1.4), получим:

$$X_{md} = 168{,}5 + \frac{(f_4 - f_3)d_3}{(f_4 - f_3) + (f_4 - f_5)} =$$

$$= 168{,}5 + \frac{(20 - 13)3}{(20 - 13) + (20 - 11)} = 168{,}5 + 1{,}312 = 169{,}81 \approx 170.$$

Соотношение между значениями среднего арифметического $\overline{X}$ медианы X_m и моды X_{md} указывает направление и степень

асимметрии распределения частот наблюдений. Если значения среднеарифметической, медианы и моды одинаковы $\bar{X} = X_m = X_{md}$, распределение симметрично. Если наибольшее значение соответствует среднеарифметической $\bar{X}$, медиана больше моды ($X_m > X_{md}$), распределение имеет положительную асимметрию. Если же наибольшее значение соответствует моде X_{md} и медиана больше среднего арифметического ($X_m > \bar{X}$), распределение имеет отрицательную асимметрию.

Перечисленные случаи показаны на рис. 1.8.

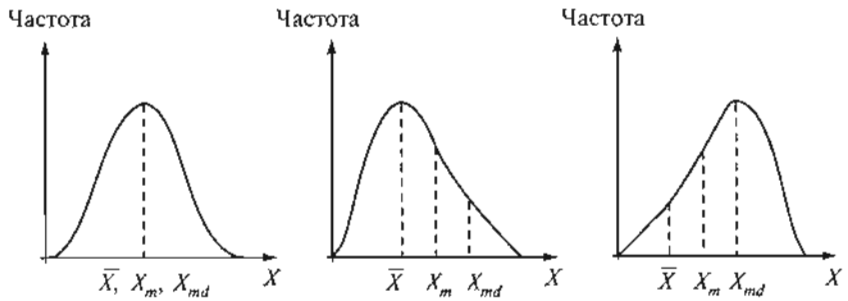

Рис. 1.8. Характеристики распределения наблюдений для различных соотношений между $\bar{X}$, X_m, X_{md}:
а — симметричное распределение ($\bar{X} = X_m = X_{md}$); *б* — положительная ассиметрия ($\bar{X} > X_m > X_{md}$); *в* — отрицательная ассиметрия ($X_{md} > X_m > \bar{X}$)

Решение вопроса о том, какая средняя наилучшая, зависит от характера распределения и предполагаемого применения этой средней. Если распределение наблюдений симметрично, можно взять любую среднюю. В случае положительной асимметрии более представительным средним является среднее арифметическое $\bar{X}$. Если же асимметрия отрицательная, предпочтение имеет мода X_{md}.

Часто интересно знать точки на оси OX, которые делят диапазон измерений на равные части.

Медиана представляет точку, которая делит накопленную частоту наблюдений так, что ее 50 % находится справа от этой точки, а 50 % слева. Иными словами, по накопленной частоте медиана делит диапазон на две половины. Квартили делят его на четыре части, децили — на 10, процентили — на 100 равных частей.

Для того чтобы найти квартили, необходимо найти три точки: Q_1, Q_2, Q_3. Тогда 25 % наблюдений слева от Q_1, 50 % слева от Q_2 и 75 % слева от Q_3 определяют квартили. Когда известна медиана $Q_2 = X_m$, следует найти только две точки: Q_1 и Q_3.

Для вычисления значений квартилей модифицируют формулу, по которой вычисляется медиана. При несгруппированных, но упорядоченных данных индекс медианы вычисляется по выражению $I_m = \frac{n}{2} + \frac{1}{2}$.

Исходя из этого индексы квартилей определяются так:

$$I_{Q1} = \frac{n}{4} + \frac{1}{2}, \quad I_{Q2} = \frac{n}{2} + \frac{1}{2}, \quad I_{Q3} = \frac{3n}{4} + \frac{1}{2}.$$

Пусть задана последовательность наблюдений 3, 5, 6, 7, 7, 8, 9, 11. Тогда $I_{Q1} = \frac{8}{4} + \frac{1}{2} = 2{,}5$ и $Q_1 = 5{,}5$. Этому индексу соответствует число 5,5.

Далее $I_{Q2} = \frac{8}{2} + \frac{1}{2} = 4{,}5$ и $Q_2 = 7$, $I_{Q3} = \frac{3 \cdot 8}{4} + \frac{1}{2} = 6{,}5$ и $Q_3 = 8{,}5$.

Аналогичным образом для упорядоченных наблюдений вычисляются децили и процентили.

При сгруппированных данных формула (1.3), используемая для вычисления медианы, преобразуется в формулу для вычисления квартилей. Предварительно находятся интервалы квартильных частот, а затем в этих интервалах осуществляется интерполяция.

$$Q_1 = a_{Q1} + \frac{\frac{n}{4} - f_{Q1-1}}{f_{Q1}} d_{Q1}, \quad Q_3 = a_{Q3} + \frac{\frac{3n}{4} - f_{Q3-1}}{f_{Q3}} d_{Q3},$$

где a_{Q1}, a_{Q3} — границы квартильных интервалов; f_{Q1-1}, f_{Q3-1} — кумулятивные частоты интервалов, непосредственно предшествующих квартильным; f_{Q1}, f_{Q3} — кумулятивные частоты квартильных интервалов; d_{Q1}, d_{Q3} — величины квартильных интервалов.

Квартили часто используют для определения межквартильного размаха и выявления выбросов наблюдений, которые для объективного их представления следует отбросить.

1.5. Числовые характеристики вариаций и формы кривой распределения

Вариация (разброс или изменчивость) данных показывает, как далеко располагаются наблюдения по отношению друг к другу, насколько расстояние между ними велико или мало.

Во многих случаях этот показатель имеет большое практическое значение. Приведем такой пример. Владелец автомобиля по мере износа шин покупает новые шины. Шины изнашиваются в разной степени.

Если все шины износились в равной степени, он сразу покупает партию (5 штук) и платит низшую цену, чем в том случае, когда он покупает шины по одной и по две. Естественно, владельца автомобиля интересуют такие шины, разброс в износе которых минимален.

Для оценки вариаций данных используют в основном такие числовые характеристики: размах вариаций, межквартильный размах, среднее отклонение, среднее квадратическое отклонение, дисперсия, коэффициент вариации.

Размах вариаций определяется по формуле

$$R = x_{max} - x_{min}, \qquad (1.5)$$

где x_{max}, x_{min} — наибольшее и наименьшее наблюдение, т. е. наибольшая и наименьшая варианта.

Например, в табл. 1.1 $x_{max} = 177$, $x_{min} = 160$, поэтому $R = 177 - 160 = 17$.

Размах вариации является грубой характеристикой изменчивости наблюдений, так как использует всего два их элемента: x_{max} и x_{min}. Достаточно измениться одному из них, и R существенно меняется. Если в приведенном примере взять $x_{max} = 180$, то R будет равно уже 20. Поэтому практическое применение этого показателя весьма ограниченно.

Однако бывают случаи, когда его использование позволяет правильно характеризовать наблюдения. Например, при оценке погрешности некоторого приближенного алгоритма Δ вычисляют среднюю погрешность Δs и размах вариации $R = \Delta_{max} - 0$.

В результате делают заключение о том, что в среднем алгоритм характеризуется погрешностью Δs, а в худшем случае может иметь погрешность Δ_{max}.

Межквартильный размах I_Q представляет собой разброс центральной половины наблюдений и вычисляется путем вычитания первого квартиля Q_1 из третьего Q_3, т. е. $I_Q = Q_3 - Q_1$.

Эта величина используется для отсева экстремальных значений наблюдений, которые искажают достоверность выборки. Практически, если наблюдения больше величины $Q_3 + 1{,}5I_Q$ или меньше величины $Q_1 + 1{,}5I_Q$, они должны быть отброшены.

Среднее отклонение основано на вычислении разности между каждой вариантой и средним арифметическим. Основанием является то соображение, что если отдельные наблюдения расположены близко друг от друга, т. е. разброс мал, то они лежат близко и от среднего значения.

Пусть имеется n наблюдений $x_1, x_2, \ldots, x_i, \ldots, x_n$ и найдено их среднее арифметическое $\bar{X}$ (формула 1.1). Назовем отклонениями от средней следующие величины $\Delta_1 = x_1 - \bar{X}$, $\Delta_2 = x_2 - \bar{X}$, ..., $\Delta_i = x_i - \bar{X}$, ..., $\Delta_n = x_n - \bar{X}$.

Тогда среднее отклонение может быть вычислено по выражению

$$\Delta_s = \frac{1}{n}\sum_{i=1}^{n}\Delta_i. \qquad (1.6)$$

Однако в связи с тем, что отклонение могут быть как положительными, так и отрицательными величинами, выражение (1.6) всегда равно нулю. Например, для ряда 4, 4, 6, 7, 9, где среднее равно 6, отклонения равны $4 - 6 = -2$, $4 - 6 = -2$, $6 - 6 = 0$, $7 - 6 = 1$, $9 - 6 = 3$, и их сумма равна нулю.

Поэтому для вычисления среднего отклонения применяют следующую формулу:

$$\Delta_s = \frac{1}{n}\sum_{i=1}^{n}|\Delta_i|, \qquad (1.7)$$

где $|\Delta_i|$, $i = 1, 2, \ldots, n$ — абсолютные величины отклонения.

В настоящее время среднее отклонение как числовая характеристика вариации используется редко.

Вместо него применяется среднее квадратическое отклонение, часто говорят стандартное отклонение.

Среднее квадратическое отклонение, как мера разброса данных, определяется через дисперсию рассеяния, интуитивно понимаемую как степень «случайности» случайной величины.

Для вычисления дисперсии используется выражение

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2. \qquad (1.8)$$

Тогда стандартное отклонение определяется так:

$$S = \sqrt{D} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2}. \qquad (1.9)$$

На практике для вычисления дисперсии и в дальнейшем стандартного отклонения выборки вместо формулы 1.8 используют в вычислительном отношении более простое выражение

$$D = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{X}^2. \qquad (1.10)$$

Оно легко выводится из выражения (1.8) путем возведения разностей $(x_i - \bar{X})$, $i = 1, 2, ..., n$, в квадрат и объединения членов.

При сгруппированных данных дисперсия согласно выражению (1.8) вычисляется с учетом того, что вместо наблюдений x_i берутся средние точки интервалов x_{i_3} и их частоты, f_i, $i = 1, 2, ...$ ..., k, а среднее арифметическое определяется по выражению (1.2). Поэтому

$$D = \frac{1}{n}\sum_{i=1}^{k}f_i \cdot (x_{i_3} - \bar{X})^2. \qquad (1.11)$$

В силу того что в выражении (1.11) берутся метки интервалов, оно дает более высокую погрешность в вычислении дисперсии, чем выражения (1.8) или (1.10).

Среднее квадратическое отклонение $S = \sqrt{D}$ очень широко используется в статистике. Для определенных случаев распределения, а именно когда оно нормальное (симметричное с нормальным эксцессом), стандартное отклонение S определяет 34 % наблюдений слева и справа от среднего арифметического $\bar{X}$. Иными словами, в интервале $[\bar{X} - S, \bar{X} + S]$ концентрируются 68 % наблюдений, а за пределами этого интервала — оставшиеся 32 % наблюдений.

Таким образом, располагая S для нормального распределения, всегда можно указать часть наблюдений, приходящихся на весь их диапазон.

Коэффициент вариации V определяется по выражению

$$V = \frac{S}{\bar{X}}. \qquad (1.12)$$

Он является относительной величиной, характеризующей степень изменчивости наблюдений. Чаще всего применяется для сравнения вариаций нескольких наборов данных.

Предположим, что в течение одного месяца средняя рыночная цена одной акции некоторой компании составляла 1500 руб. при стандартном отклонении $S = 150$ руб. Средняя рыночная цена одной акции другой компании была равна 500 руб. при стандартном отклонении $S = 50$ руб.

Необходимо установить, цена какой акции более изменчива.

На первый взгляд можно сказать, что поскольку 150 руб. > > 50 руб., т. е. стандартное отклонение $S = 150 > S = 50$, цена первой акции более изменчива. Однако в действительности это не так, в чем можно убедиться, сравнивая соответствующие коэффициенты вариации. Действительно, $V_1 = \frac{150}{1500} = 0{,}1$ и $V_2 = \frac{50}{500} = 0{,}1$.

Поэтому по отношению к соответствующим средним ценам колебание цен обеих акций одинаково.

Раньше асимметрия и эксцесс определялись по геометрической форме кривой распределения. Однако они могут быть установлены и по соответствующим значениям специальных числовых показателей.

Для этого в статистике используются специальные величины, называемые моментами распределения относительно средней арифметической.

Термином «момент» называют сумму степеней отклонений от средней арифметической. Первый момент m_1 вычисляется по выражению (1.6) и всегда равен нулю. Второй момент m_2 вычисляется по формуле (1.10) и определяется как дисперсия выборки. Третий момент m_3 вычисляется по формуле

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}\Delta_i^3, \qquad (1.13)$$

т. е. представляет собой сумму кубов отклонений от среднего арифметического. Для вычисления четвертого момента m_4 служит выражение

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}\Delta_i^4. \qquad (1.14)$$

Третий момент m_3 используют для формирования показателя асимметрии распределения $a_3 = \frac{m_3}{S_3}$. Если $a_3 = 0$, распределение симметрично. Если $a_3 < 0$, то распределение характеризуется отрицательной асимметрией, если $a_3 > 0$ — положительной асимметрией.

Таким образом, для определения характера асимметрии рассматривают только знак показателя a_3.

Для определения вида эксцесса вводят показатель $a_4 = \frac{m_4}{S^4} - 3$. В том случае, когда $a_4 = 0$, распределение характеризуется нормальным эксцессом. Если $a_4 < 0$, эксцесс ниже нормального, если $a_4 > 0$ — выше нормального.

Таким образом, как упоминалось, по соотношениям между средним арифметическим $\overline{X}$, медианой X_m и модой X_{md} можно установить вид асимметрии кривой распределения. С другой стороны, вид распределения можно установить и по знакам коэффициентов a_3, a_4.

1.6. Компьютерные технологии описательной статистики

Компьютерные технологии представляют собой процедуры решения задач обработки информации. Они предназначены для решения задач как описательной, так и аналитической статистики.

Практически эти технологии реализованы либо в виде профессиональных пакетов прикладных программ, таких как Statgrafics, Statistica, Stadia, либо как комплексы программ, входящих в различные математические пакеты, например Mathcad, Mathlab, Maple.

Табличные процессоры Microsoft Excel также включают комплекс программ, предназначенный для решения задач статистики. Он называется **Пакет анализа** и представляет собой одну из программных надстроек (дополнений) к Excel.

Роль табличного процессора Excel при практическом использовании **Пакета анализа** для решения тех или иных задач статистики сводится в основном к выполнению следующих процедур:

1) занесение в книгу Excel исходной таблицы наблюдений;
2) сохранение этих наблюдений и промежуточных результатов расчетов на жестком диске в виде файлов;
3) чтение данных с диска в оперативную память компьютера;
4) печать необходимых результатов расчетов;
5) использование функций;
6) построение графиков.

Поэтому, чтобы иметь возможность применять процедуры **Пакета анализа** на практике, статистик должен уметь выполнять перечисленные действия в активном режиме Excel.

Инициация **Пакета анализа** осуществляется после инициации табличного процессора Excel. Для этого в строке меню первой книги, которая появляется на экране в результате запуска Excel, необходимо щелкнуть мышью **Сервис**.

Далее в раскрывшемся меню **Сервис** необходимо щелкнуть пункт **Анализ данных** и сразу попасть в окно **Надстройки**, либо если такого пункта нет, щелкнуть пункт меню **Надстройки** и попасть в окно перечня **Надстроек**. После этого в окне списка **Надстроек** необходимо щелкнуть пункт **Пакет анализа**.

В результате выполнения этих действий в раскрывшемся окне появится список реализованных в надстройке **Анализ данных** методов статистической обработки данных. Он включает следующие пункты.

1. Гистограмма.
2. Выборка.
3. Описательная статистика.
4. Ранг и персентиль.
5. Генерация случайных чисел.
6. Двухвыборочный z-тест для средних.
7. Двухвыборочный t-тест с одинаковыми дисперсиями.
8. Двухвыборочный t-тест с различными дисперсиями.
9. Двухвыборочный F-тест для дисперсий.

10. Парный двухвыборочный *t*-тест для средних.
11. Однофакторный дисперсионный анализ.
12. Двухфакторный дисперсионный анализ без повторений.
13. Двухфакторный дисперсионный анализ с повторениями.
14. Ковариация.
15. Корреляция.
16. Регрессия.
17. Скользящее среднее.
18. Экспоненциальное сглаживание.
19. Анализ Фурье.

Для инициации любого пункта необходимо выделить его мышью и щелкнуть кнопку **ОК**. В результате на экране появится диалоговое окно с элементами управления — полями ввода, переключателями, раскрывающимися списками и кнопками. Набор этих элементов для каждого пункта специфичен.

После заполнения необходимых полей и нажатия кнопки **ОК** на экран будут выведены результаты выполнения инициированного пункта.

Пакет прикладных программ Excel содержит большое количество программных компонент, предназначенных для вычисления значений различных функций. Функции классифицированы по тематическому признаку: финансовые, математические, статистические и др.

Программная компонента Excel, управляющая работой пользователя с функциями, называется **Мастер функций**.

Инициация этой компоненты осуществляется из пункта главного меню **Вставка**, для чего необходимо щелкнуть этот пункт, или из строки инструментов, для чего необходимо щелкнуть значок f_x. Далее в появившемся свисающем меню необходимо щелкнуть пункт **Категория**.

В результате на экране появится окно мастера с перечнем десяти категорий функций. Для работы с функциями определенной категории, в нашем случае статистическими, необходимо выделить эту категорию и щелкнуть по ней мышью.

Excel содержит 78 статистических функций. Безусловно, имеются функции, определяющие рассмотренные выше числовые характеристики описательной статистики.

Для того чтобы реализовать ту или иную функцию, необходимо ее выделить и щелкнуть мышью кнопку **ОК**. В появившемся окне ввести поля, требуемые аргументы и снова щелкнуть кнопку **ОК**.

По каждой функции и в окне **Мастера функций**, и в окне **Аргументы функции** даются подробные пояснения. При необходимости для уточнения требуемых сведений можно использовать справку по функции: щелкнуть по надписи в окне **Справка по функции**.

Статистические функции могут использоваться как автономно, так и в различных режимах пакета **Анализ данных**.

В меню надстройки **Анализ данных** для реализации методов описательной статистики предусмотрено два пункта: **Гистограмма** и **Описательная статистика**.

Пункт **Гистограмма** позволяет построить ряд распределения частот исходной выборки, гистограмму и кривую накопленных частот.

Для построения полигона необходимо использовать программу **Мастер диаграмм**, которая вызывается из пункта главного меню Excel **Вставка**. После нажатия этой клавиши в свисающем меню необходимо выделить пункт **Диаграмма** и щелкнуть по нему.

Далее в открывшемся меню **Мастера диаграмм** следует найти пункт **График**, щелкнуть по нему мышью и для вывода полигона щелкнуть кнопку **Готово**.

Рассмотрим на примере (данные табл. 1.1) построение ряда распределения, гистограмму, полигон, а также рассчитаем статистику выборки.

Прежде всего необходимо средствами Excel **ввести** данные табл. 1.1 в память компьютера. Для этого в ячейки **(A1**:D1) поместим 50 чисел этой таблицы. Кроме этого, в ячейки **(E1**:E7) занесем границы интервалов, а в ячейки (I1:I8) — метки (середины интервалов), которые были найдены раньше в процессе демонстрации ручного построения ряда распределения гистограммы и полигона.

Введенные данные обязательно сохраним на жестком диске простейшим способом: из пункта основного меню **Файл** выполним команду **Сохранить**. В результате файл будет сохранен на диске C: в папке **Мои документы** под именем Книга1.

После этого щелкнем пункт меню **Сервис** и в открывшемся свисающем меню — пункт **Анализ данных**. В появившемся диалоговом окне найдем пункт **Гистограмма** и нажмем кнопку **ОК**.

В результате этих действий на экране появится диалоговое окно **Гистограмма** с необходимыми элементами управления: входной интервал, интервал карманов, метки и параметры вывода.

Во входной интервал поместим диапазон ячеек B1:B50. Заполнение интервала карманов не обязательно, хотя в него можно поместить числа, определяющие границы интервалов: 159,5 162,5, ...

Далее щелкнем в окошках метки: *новый рабочий лист*, *интервальный процент*, *вывод графика* и *кнопку* **ОК**. В результате на экране появятся ряд распределения и гистограмма.

Для построения полигона обратимся к **Мастеру диаграмм**. Щелкнем пункт меню **Вставка**, затем пункт **Диаграмма**, затем пункт **График**. После этого щелкнем пункт **Далее**, в результате чего на экран будет выведен полигон.

Представление крупным планом гистограммы и полигона и их оформления (заголовки), надписи по осям осуществляются из окна **Мастера диаграмм**. Нажатие кнопки **Далее** показывает, какие действия необходимо последовательно выполнять. В последнем окне на вопрос, на каком листе поместить диаграмму, необходимо щелкнуть **Отдельном**, а затем нажать клавишу **Готово**.

Вычисление статистик осуществляется из окна надстройки **Анализ данных**. В этом окне необходимо щелкнуть пункт **Описательная статистика** и нажать кнопку **ОК**.

В результате раскроется диалоговое окно, содержащее необходимые элементы управления. Далее во входной интервал необходимо занести диапазон ячеек B1:B50, щелкнуть в окошках *метки*, *итоговая ситуация*, *уровень надежности*, *k-й наименьший* и *k-й наибольший* и *нажать кнопку* **ОК**. В результате на экране появятся расчетные данные в виде табл. 1.6.

Таблица 1.6. **Статистики роста студентов**

Среднее	169,76
Ст. ошибка	0,446
Медиана	170
Мода	170
Ст. отклонение	3,153

Окончание табл. 1.6

Дисперсия	9,941
Эксцесс	1,003
Асимметр	–0,319
Интервал	17
Минимум	160
Максимум	177
Сумма	8321
Счет	50
Наибольшее	177
Наименьшее	160
Уровень надежности	0,896

Как следует из этой таблицы, распределение роста студентов практически симметричное: среднее арифметическое, мода и медиана почти равны. Оно имеет незначительную отрицательную асимметрию и эксцесс выше нормального. Это означает, что распределение значительно концентрированно вблизи среднего арифметического, о чем свидетельствуют также малая величина стандартного отклонения.

Контрольные вопросы

1. Дайте толкование случайному и детерминированному явлениям. Приведите примеры таких явлений.
2. Определите понятие случайного события. Как понимать исход события и наблюдение?
3. Какие события называют достоверными, невозможными, равновозможными, несовместимыми?
4. Как определяются частота и частость события?

5. В каких пределах лежит частость события?
6. Дайте толкование понятию «случайная величина». Приведите примеры таких величин.
7. Какие типы случайных величин вам известны? Приведите примеры таких величин.
8. Определите понятия сплошного и выборочного изучения некоторого явления, генеральной и выборочной совокупности наблюдений.
9. Как называются числовые характеристики генеральной совокупности наблюдений и выборки?
10. Чем объясняется изучение случайных событий только частью наблюдений?
11. Какие задачи ставят перед собой описательная и аналитическая статистика?
12. Что представляет собой ряд распределения наблюдений и зачем строят ряды?
13. Как определяется количество интервалов ряда распределения?
14. Как распределяются нижние и верхние границы ряда распределений, метки?
15. Можно ли построить ряд распределения для дискретных величин?
16. Существуют ли графические способы представления рядов распределения?
17. Как строятся гистограмма, полигон и кривая накопленных частот?
18. Что представляют собой симметричная и асимметричная кривые распределения частот?
19. Как классифицируются кривые по характеру асимметрии?
20. Что характеризует эксцесс кривой?
21. Как классифицируются кривые по характеру эксцесса?
22. В чем смысл понятия «центр распределения», какие числовые характеристики используются для его определения?

23. Как вычислить среднее арифметическое, когда исходные данные представлены таблицей, рядом распределения?
24. Что представляет собой медиана и как она вычисляется для исходных данных, представленных таблицей и рядом распределения?
25. Что представляет собой мода? Есть ли последовательности, не содержащие моду?
26. Как определяется мода для несгруппированных данных и ряда распределения?
27. Можно ли по числовым соотношениям между средним арифметическим, медианой и модой судить о симметричности или асимметричности кривой распределения?
28. Что представляют собой квартили, децили, процентили?
29. Какие числовые показатели используются для характеристики разброса наблюдений?
30. Как определяется размах вариации? В чем недостаток этой характеристики?
31. Как определяется среднее отклонение?
32. Как определяется среднее квадратическое отклонение?
33. Как можно вычислить дисперсию?
34. Как определяется коэффициент вариации? Что он характеризует?
35. Истолкуйте термин «момент».
36. Какие моменты используются для определения асимметрии и эксцесса?
37. Как называется программная надстройка Microsoft Excel, предназначенная для решения задач статистики?
38. Как инициировать надстройку?
39. Какие пункты меню окна **Анализ данных** используются для решения задач описательной статистики?
40. Как инициировать любой пункт меню?
41. Как пользоваться статистическими функциями?

42. Как построить рад распределения и гистограмму?
43. Как построить полигон?
44. Как вычисляются статистики?
45. Можно ли, имея некоторый ряд распределения, построить гистограмму и полигон?

Задачи

1. Постройте ряд распределения для следующей таблицы исходных данных.

Таблица 1.7. **Случайные наблюдения**

32	26	16	44	28	40	30	31	17	30
37	32	42	31	36	49	35	21	25	40
27	25	33	34	27	35	43	28	19	20

При этом определите количество интервалов по формуле Штюргеса и окончательное количество интервалов. Установите границы интервалов и средние точки.

2. Используя ряд распределения, постройте гистограмму, полигон и кривую накопленных частот.
3. По гистограмме и полигону определите примерную форму кривой распределения наблюдений.
4. Вычислите среднее арифметическое по исходной таблице данных и по ряду распределения.
5. Найдите медиану и моду.
6. Установите соотношение между средним арифметическим медианы и моды.
7. Дайте на этом основании заключение о форме кривой.
8. Найдите такие статистики выборки: размах вариации, межквартильный размах, дисперсию, стандартное отклонение, коэффициент для определения симметрии и эксцесса.

9. Перечисленные пункты задания выполните на компьютере.
10. Постройте гистограмму и полигон на компьютере, используя для этого данные ряда распределения, построенного вручную.
11. Объясните различия в гистограмме и полигоне, полученных ручным способом и на компьютере.
12. Применяя компьютерную технологию, выполните статистический анализ данных, представленных в табл. 1.8.

Таблица 1.8. **Пробег 60 шин нового типа, тыс. км**

40,1	41,9	47,9	42,8	43,3	47,5	43,9	42,6
45,8	50,2	46,1	40,7	46,7	47,2	43,6	50,8
44,1	45,0	48,9	46,4	46,3	42,9	46,9	42,3
52,0	41,3	47,7	40,4	45,5	42,1	48,2	47,7

39,1	43,9	46,9	46,7	49,1	37,4	43,9
40,6	44,7	45,2	48,8	51,2	46,9	43,6
44,5	43,1	47,0	44,2	44,4	43,4	41,8
43,7	48,3	42,6	49,8	45,5	41,5	44,8

АНАЛИТИЧЕСКАЯ СТАТИСТИКА

Глава 2
ВЕРОЯТНОСТЬ И СТАТИСТИКА

2.1. Элементы теории вероятностей

В процессе выработки статистического заключения на основе выборки, т. е. перехода от частного к общему, а также принятия различных управленческих решений, невозможно исключить риск. Практика показывает, что заключения в какой-то мере могут оказаться ошибочными, а исходы управленческих решений не те, которые ожидались. Если была бы числовая мера оценки риска, многих ошибок удалось избежать. При высоком значении риска решения бы не принимались, а предварительные заключения — дорабатывались.

Оценить степень риска в статистике помогает применение теории вероятностей — раздела математики, в котором изучаются количественные связи в области случайных событий, величин и процессов.

Вероятностью случайного события A называют числовую оценку объективной возможности появления этого события. Для того чтобы определить интервал, которому принадлежит такая оценка, условились, что вероятность достоверного события равна 1, а вероятность невозможного — 0.

Таким образом, вероятность события A (обычно обозначают $P(A)$) лежит в интервале $0 \le P(A) \le 1$.

Чем достовернее событие A, т. е. более правдоподобное, тем больше значение $P(A)$. И наоборот, чем менее правдоподобное событие A, тем меньше значение $P(A)$.

Часто говорят, событие A более вероятно, чем событие B, имея в виду, что вероятность $P(A)$ больше вероятности $P(B)$. Например, при бросании игральной кости более вероятное событие — выпадение четного числа очков, чем выпадение шести очков.

Такой вывод обосновывается тем, что четное число очков будет выпадать чаще, чем появляться шесть очков, так как четных чисел больше — 2, 4, 6, чем число очков 6. При бросании монеты вероятности выпадения орла или решки одинаковы, так как эти события равнозначны.

Существует два способа вычисления вероятности некоторого события A. Первый способ классический, который может быть назван аналитическим, второй — эмпирический, часто именуемый частотным. Оба способа считаются объективными, а вероятность, найденная одним из них, — объективной.

При классическом подходе к вычислению вероятности делают предположение о том, что наблюдаемое явление или результаты опыта имеют n исходов, т. е. могут быть представлены n событиями, которые несовместные (взаимоисключающие) и равновозможные.

Тогда вероятность появления некоторого события A может быть вычислена по выражению

$$P(A) = \frac{n_A}{n}, \qquad (2.1)$$

где n_A — число случаев появления события A из всего возможного их числа n, часто говорят, число случаев, благоприятных событию A.

Например, при **подбрасывании монеты** общее число событий равно двум: выпадение орла или решки. Эти события несовместные и равновозможные. Числа выпадения орла n_o или решки n_r равны 1. Поэтому согласно формуле (2.1) вероятность событий «выпадение орла» или «выпадение решки» равна $\frac{1}{2}$.

При вытягивании одной карты из колоды 52 карт, вероятность появления любой карты любой масти равна $\frac{1}{52}$, потому что всего исходов 52, а благоприятный исход (любая карта) один.

Однако вероятность появления, например, туза равна $\frac{4}{52} = \frac{1}{13}$, так как благоприятных исходов 4: в колоде — 4 туза.

В некоторых задачах общее число исходов может быть очень велико и требовать специального подсчета.

В качестве примера рассмотрим известную задачу **о спортивной лотерее**: угадать шесть номеров из 49. Подсчитаем вероятность этого события. Для этого определим общее число исходов. Очевидно, что оно равно числу комбинаций шести номеров из 49.

Как известно, это число комбинаций равно числу сочетаний C_{49}^{6}, которое определяется по выражению

$$C_{49}^{6} = \frac{49!}{(49-6)!\cdot 6!} = \frac{49!}{43!\cdot 6!} = \frac{44\cdot 45\cdot 46\cdot 47\cdot 48\cdot 49}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6} = 13\,983\,616.$$

Таким образом, полученное число сочетаний — число возможных исходов. Благоприятный исход (угадывание шести очков) — один. Поэтому вероятность этого исхода $P = \frac{1}{13\,983\,616} =$ $= 0{,}00000007$. По всей видимости, играть в такую лотерею не имеет смысла.

Классический способ вычисления вероятности ограничен весьма жесткими условиями, которые в большинстве практических случаев не выполняются. Чаще всего исходы не являются равновозможными. Поэтому применение такого способа вычисления вероятности не всегда оказывается возможным.

При частотном способе вычисления вероятности некоторого события A полагают, что вероятность этого события равна относительной частоте появления A при наблюдении очень большого числа событий. Иными словами,

$$P^{*}(A) = W_n(A) = \frac{n_A}{n} \quad \text{при} \quad n \to \infty, \qquad (2.2)$$

где n_A — число появлений события A в общем числе наблюдений n;

∞ — символ бесконечности.

Частотный способ определения вероятности основан на здравом смысле — более вероятные события происходят чаще маловероятных — и подтверждается практическим опытом.

В свое время многие ученые пытались доказать правомерность этого подхода: многократно подбрасывали монеты и бросали игральные кости. Например, известный статистик К. Пирсон подбросил монету 24 000 раз и получил 12 012 выпадений орла.

Это дает частость $W_n(A)$, примерно равную $\frac{1}{2}$. Такова же и вероятность выпадения орла или решки, получаемая классическим методом.

Что же является причиной приближения относительной частоты к вероятности события при увеличении n? В основе лежит закон устойчивости частот, наблюдаемых в массовых случайных явлениях, при проведении большого числа однородных опытов или природных наблюдений. Чем больше этих опытов, тем частота события становится все менее случайной, стабилизируется и приближается к постоянной. Эта постоянная и является вероятностью события.

Впервые теоретически это положение доказал известный математик XVI в. Я. Бернулли. В современной теории вероятности оно представляет собой одну из форм закона больших чисел, закона об устойчивости определенных характеристик массовых случайных явлений.

Недостаток частотного способа вычисления вероятности некоторого события A состоит в том, что данный способ дает приближенное ее значение, которое можно уточнить, увеличивая число опытов.

Теория вероятностей предусматривает возможность выполнения над вероятностями нескольких событий двух операций: сложения и умножения вероятностей. Для классического способа определения вероятности эти операции легко доказуемы, для частотного — применяются как аксиомы (на веру).

Зная вероятности простейших событий, данные операции позволяют вычислить на их основе вероятности событий более сложных, которые часто другими способами не поддаются определению.

Правило сложения вероятностей формулируется для несовместных (взаимоисключающих) и совместных событий. Если события A, B несовместны, то вероятность того, что произойдет одно из них, неважно какое, равна сумме вероятностей событий A и B.

В виде формулы это записывается так:

$$P(A \text{ или } B) = P(A) + P(B). \qquad (2.3)$$

Например, при подбрасывании монеты вероятности появления орла — событие A, появление решки — событие B — равны $\frac{1}{2}$. Поэтому сумма вероятностей $P(A$ или $B)$ в этом случае равна 1. Более того, события A и B не только несовместные, но и противоположные.

Следовательно, если A — некоторое событие, а $\bar{A}$ — событие, ему противоположное, то $P(A) = 1 - P(\bar{A})$. Это дает возможность вычислять вероятности через противоположные события. Если вероятность интересующего нас события A вычислить трудно, а вероятность противоположного ему — легко, то $P(A) = 1 - P(\bar{A})$.

Правило сложения вероятностей обобщается на любое число событий. Вероятность того, что произойдет любое из нескольких несовместных событий, равна сумме вероятностей этих событий, т. е.

$$P(A \text{ или } B... \text{ или } X) = P(A) + P(B) + ... + P(X). \qquad (2.4)$$

В том случае, когда рассматриваемые события A, B — совместные, например вытягивание короля или туза из колоды 52 карт при условии, что берется одна карта, вероятность события A или B определяется так:

$$P(A \text{ или } B) = P(A) + P(B) - P(A, B), \qquad (2.5)$$

где $P(A, B)$ — вероятность появления как A, так и B.

Отсюда, в частности, следует, что формула (2.3), согласно которой определяется вероятность события A или B, является частным случаем формулы (2.5), так как в формуле (2.3) вероятность $P(A, B) = 0$.

Так, вероятность одновременного вытягивания туза или короля из колоды с 52 картами равна $\frac{4}{52} + \frac{4}{52} = \frac{2}{13}$ в связи с тем, что эти события несовместные.

Вероятность вытягивания короля и карты, например, пиковой масти равна $\frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{4}{13}$, так как эти события совместные: $\frac{13}{52}$ – вероятность вытягивания любой пиковой масти, поскольку в колоде по 13 карт каждой масти, а $\frac{1}{52}$ — вероятность вытягивания как туза, так и любой карты.

Правило умножения вероятностей формулируется для зависимых и независимых событий.

Если вероятность одного из событий зависит от результата исхода другого события, такие события называются *зависимыми*. В противном случае они независимы. Например, при многократном бросании монеты вероятность выпадения орла не зависит от того, что выпало в предшествующем бросании. Она равна $\frac{1}{2}$. То же можно сказать и о решке.

Однако если из ящика с 7 шарами, 3 из которых белые, а 4 черные, вынут белый шар, то вероятность снова вынуть белый шар зависит от предшествующего результата, так как эти события зависимые.

Подсчитаем вероятность в очередной раз вынуть белый шар. Вероятность первого исхода — вынуть белый шар — равна $\frac{3}{7}$, так как всего исходов 7, а благоприятствующих рассматриваемому событию — 3 (классическая формула 2.1). Вероятность во второй раз вынуть белый шар находится по той же классической формуле и равна $\frac{2}{6} = \frac{1}{3}$.

Если события A, B независимы, вероятность их совместного исхода определяется по выражению

$$P(A \text{ и } B) = P(A) \cdot P(B). \qquad (2.6)$$

Если же события A, B зависимы, то вероятность совмещения этих событий вычисляется так:

$$P(A \text{ и } B) = P(A)\ P(B/A) = P(B)\ P(A/B), \qquad (2.7)$$

где $P(B/A)$, $P(A/B)$ — условные вероятности, вычисленные при условии, что произошло событие A либо B.

Это и есть правила умножения вероятностей. В примере с семью шарами мы уже вычислили условную вероятность событий: A — первый шар белый. B — второй шар белый. Поэтому

вероятность совмещения этих событий — оба шара белые — определяется по формуле (2.7). Таким образом, $P(A \text{ и } B) = \frac{3}{7} \cdot \frac{1}{3} = \frac{1}{7}$.

В том случае, когда рассматривается множество независимых событий A, B, ..., X, вероятность их совместного исхода определяется по выражению

$$P(A \text{ и } B \text{ и } \ldots \text{ и } X) = P(A) \cdot P(B) \cdot \ldots \cdot P(X). \qquad (2.8)$$

Если же события зависимы, вероятность их вычисляется так: определяется вероятность одного события, умножается на условную вероятность другого в предположении, что первое событие произошло, затем умножается на условную вероятность третьего события в предположении, что первых два события произошли, и т. д.

Рассмотрим пример, который демонстрирует применение правил сложения и умножения вероятностей. Согласно ему решаются многие практические задачи.

Производится n независимых опытов. В каждом опыте вероятность исхода некоторого события A равна $P(A)$. Найти вероятность того, что в n опытах это событие появится хотя бы один раз.

Решается эта задача переходом к противоположному событию $\bar{A}$. Так как вероятность исхода события A равна $P(A)$, то по правилу сложения вероятностей независимых событий вероятность противоположного события $\bar{A} = 1 - P(A)$.

По правилу умножения вероятностей n независимых событий

$$P(\bar{A}) = \underbrace{(1 - P(A)) \cdot ((1 - P(A)) \cdots (1 - P(A))}_{n \text{ раз}} = (1 - P(A))^n,$$

т. е. это вероятность непоявления события A. Переходя от вероятности $P(\bar{A})$ к вероятности $P(A)$, получаем $P(A) = 1 - (1 - P)^n$, т. е. вероятность появления хотя бы одного события A.

Предположим, вероятность P обнаружения летательного объекта при одном цикле кругового обзора радиолокационной станции равна 0,1. Найти вероятность того, что после 10 циклов обзора аппарат будет обнаружен.

Вероятность необнаружения летательного аппарата в одном цикле обзора равна 1 – 0,1 = 0,9. Вероятность необнаружения этого аппарата при 10 циклах обзора равна $0{,}9^{10} = 0{,}348$. Вероятность обнаружения аппарата в 10 циклах обзора равна 1 – 0,348 = 0,652.

Рассмотренные классический и эмпирический подходы к вычислению вероятности опирались на основную мысль: числовое значение вероятности порождается частотой появления рассматриваемого события в некоторой последовательности наблюдений.

Как же поступают в том случае, когда указанные подходы не применимы? Например, когда нужно предсказать появление того или иного исхода некоторого единичного события?

В этом случае вероятность толкуется как степень уверенности лица в том, что благоприятный исход произойдет. Такой подход определения вероятности часто называют субъективным, он широко используется в управленческих решениях, когда не представляется возможным определить частоту идентичных событий.

2.2. Распределения вероятностей

При построении рядов распределения частот весь диапазон наблюдений разбивался на интервалы и каждому из них сопоставлялась частота — число наблюдений, попадавших в этот интервал. Далее вычислялись относительные частоты, в результате чего получали распределение частот.

В связи с тем что при увеличении числа наблюдений согласно закону больших чисел частота приближается к вероятности, правомерно заменить частоты вероятностями и получить распределение вероятностей, в котором каждому наблюдению будет соответствовать своя вероятность. Это распределение принято называть законом распределения вероятности, а термин «распределение» означает, что вероятность как-то распределена между 0 и 1.

Поскольку вероятность случайного события, степень уверенности в реальности того, что это событие произойдет, закон распределения вероятности, представляющий весь диапазон наблюдений, дает возможность оценить риски, возникающие в про-

цессе статистических заключений и принятии управленческих решений в значениях вероятности.

Существует много теоретических законов распределения вероятности: нормальный, равновероятный, биномиальный, геометрический, закон Пуассона и др. Все эти законы используются для решения тех или иных задач теории вероятности и математической статистики.

Ниже рассмотрены три наиболее часто применяемых в аналитической статистике закона: биномиальный, закон Пуассона и нормальный закон. Первые два закона описывают распределение вероятности дискретной, третий — непрерывной случайной величины.

Предположим, осуществляется подбрасывание трех одинаковых монет и ведутся подсчеты количеств выпадения орла. Возможные четыре исхода каждого подбрасывания: 0 выпадений, 1 выпадение, 2 выпадения, 3 выпадения орла.

При определенном количестве подбрасывания монет можно построить ряд распределения частот, представленный в табл. 2.1 для двенадцати подбрасываний.

Таблица 2.1. **Ряд распределения частот выпадений орла**

Число выпадений орла	Частота	Относительная частота	Вероятность
1	2	1/6	1/8
2	5	5/12	3/8
3	4	1/3	3/8
3	1	1/12	1/8

Если количество подбрасываний существенно увеличить, то относительные частоты согласно закону больших чисел окажутся близки к вероятностям чисел выпадений орла. Вместе с тем эти вероятности могут быть вычислены априори (до опыта) по специальному правилу.

Было установлено, что они могут быть найдены как коэффициенты разложения бинома Ньютона $(a + b)^n$. Отсюда и происходит название закона — биномиальное распределение.

Рассмотрим, как вычислить вероятности. Прежде всего символам a и b приписываются значения вероятностей, которые должны быть известны для каждой задачи. В нашем случае есть

три монеты, подбрасывание каждой из которых имеет два исхода: орел или решка. Вероятности этих событий $\frac{1}{2}$.

Таким образом, $a = \frac{1}{2}$, $b = \frac{1}{2}$, а их сумма равна 1. Дальше необходимо определить значение степени бинома n.

Показатель степени принимается равным числу объектов, участвующих в эксперименте. Если подбрасывается одна монета, $(a + b)^1$, если подбрасывается три монеты, $n = 3$.

Таким образом, получаем: $(a + b)^3 = a^3 + 3a^2b + 3b^2\ a + {}_b{}^3$.

Подставляя вместо a, b их вероятности, будем иметь:

$$(1/2)^3 + 3(1/2)^2(1/2) + 3(1/2)^2(1/2)^3 + (1/2)^3 =$$

$$1/8 + 3/8 + 3/8 + 1/8 = 1.$$

На этом основании вероятности 0, 1, 2, 3 исходов соответственно равны 1/8, 3/8, 3/8, 1/8.

Эти вероятности занесены в табл. 2.1, и, следовательно, к ним должны стремиться частоты при увеличении числа подбрасываний монет.

Теперь рассмотрим более реальный пример. Некоторое предприятие изготавливает электродвигатели и, прежде чем выставлять их на продажу, определяет, имеет ли двигатель дефект. При этом предшествующим опытом установлено, что 10 % двигателей имеют какие-либо неисправности. Контроль осуществляется по партиям, включающим три двигателя.

Требуется построить закон распределения вероятностей обнаружения различного числа дефектных двигателей.

Для каждой партии из трех двигателей при контроле имеем исходы, представленные в первой колонке табл. 2.2.

Таблица 2.2. **Биномиальное распределение вероятностей**

Число дефектных двигателей	Вероятность
0	0,729
1	0,243
2	0,027
3	0,001

Во второй колонке этой таблицы представлены вероятности, полученные на основании разложения бинома $(a + b)^3$, с учетом того, что вероятность обнаружения дефектного двигателя b = = 10 % = 0,1, следовательно, бездефектного — 0,9.

Подставляя в разложение $a^3 + 3a^2b + 3b^2a + b^3$ значения $a = 0{,}9$ и $b = 0{,}1$, получим:

$$(0{,}9)^3 + 3(0{,}9)^2 \cdot 0{,}1 + 3(0{,}1)^2 \cdot 0{,}9 + (0{,}1)^3 =$$

$$= 0{,}729 + 0{,}243 + 0{,}027 + 0{,}001 = 1.$$

По этому распределению можно судить, что вероятность в партии бездефектных двигателей равна 0,729 = 72,9 %, с одним дефектным двигателем — 0,243 = 24,3 %, с двумя дефектными двигателями — 0,027 = 2,7 %, тремя дефектными двигателями — 0,001, т. е. 0,1 %.

При рассмотрении биномиального распределения благоприятный исход принято называть успехом, а неблагоприятный — неудачей. Соответственно именуются и вероятности: как вероятность успеха и неудачи.

В том случае, когда эти вероятности равны, биномиальное распределение симметрично. В противном случае оно асимметрично. Применительно к рассмотренной задаче эти распределения показаны на рис. 2.1.

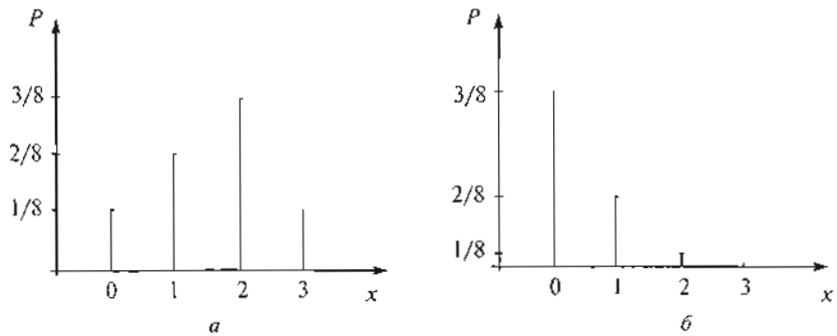

Рис. 2.1. Биномиальное распределение:
а — вероятность благоприятного исхода $P = 0{,}2$; *б* — вероятность благоприятного исхода $P = 0.9$

Биномиальное распределение дает решение задачи, когда проводятся многократно повторяемые испытания с двумя исходами и неизменной вероятностью этих исходов: P — успех, $(1 - P)$ — неудача.

Вероятность того, что появится x успешных исходов в n испытаниях, определяется по формуле Бернулли

$$P(x) = C_n^x P^x (1 - P)^{n-x}, \quad x = 0, 1, 2, \ldots, n, \qquad (2.9)$$

где C_n^x — число сочетаний элементов из n по x — коэффициент, при соответствующем члене разложения бинома Ньютона.

Например, если x — все успешные исходы при трех бросаниях монет ($n = 3$), то $P(x) = C_3^3 \cdot P^3 (1 - P)^{3-3} = P^3$ — первый член разложения бинома $(a + b)^3$, так как $C_3^3 = 1$, $(1 - P)^0 = 1$.

Таким образом, для того чтобы подсчитать вероятность успешных исходов в n независимых испытаниях, следует использовать формулу (2.9).

Например, пусть осуществляется подбрасывание симметричной монеты. Вероятность успеха $P = 1/2$, неудачи тоже 1/2. Спрашивается, какая вероятность выпадения 7 орлов (успехов) при 10 подбрасываниях монеты? Согласно выражению (2.9) имеем:

$$P(7) = C_{10}^7 \cdot \left(\frac{1}{2}\right)^7 \cdot \left(\frac{1}{2}\right)^3 = C_{10}^7 \cdot \left(\frac{1}{2}\right)^{10} = \frac{10!}{(10 - 7)!\,7!} \cdot \left(\frac{1}{2}\right)^{10} =$$

$$= \frac{8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3} \cdot \left(\frac{1}{2}\right)^{10} = \frac{4 \cdot 3 \cdot 10}{1024} = \frac{120}{1024} = 0{,}172.$$

Среднее значение биномиального распределения вероятностей $\bar{X} = n \cdot P$, стандартное отклонение $S = \sqrt{n \cdot P(1 - P)}$, где n — число испытаний; P — вероятность успеха.

Распределение Пуассона используется для вычисления вероятности того, что некоторое событие произойдет в течение заданного интервала времени, например вероятности, что в течение следующего часа в супермаркет войдет десять покупателей. Случайной переменной распределения Пуассона в данном случае будет фактическое число покупателей, вошедших в магазин в течение очередного часа.

Среднее значение распределения Пуассона — это среднее количество реализаций за ряд интервалов, которые происходят в

заданный период времени. Причем число реализаций в каждом интервале не зависит от числа реализаций в других интервалах.

Таким образом, распределение Пуассона характеризует эксперимент или явление, в котором подсчитывается число рассматриваемых событий, происходящих за определенный период времени. Примерами таких событий могут быть количество машин, прибывающих на некоторую бензозаправку в течение часа, количество студентов, отсутствующих на лекции по статистике, читаемой по средам, и т. д.

Если некоторая случайная величина x имеет распределение Пуассона, то вероятность реализации ее за определенный промежуток времени определяется по выражению $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, где λ — среднее количество реализаций x за определенный промежуток времени, а $e = 2{,}718$ — константа Эйлера.

Предположим, известно, что на определенном перекрестке среднее количество аварий $\lambda = 1{,}8$. Подсчитаем вероятность того, что в очередном месяце произойдет три или большее число аварий. Иными словами, необходимо найти вероятность $P(x \geq 3)$. Так как

$$P(x \geq 3) = P(x = 3) + P(x = 3) + P(x = 4) + \ldots + P(x = \infty),$$

подсчитать сумму этих вероятностей, пользуясь уравнением Пуассона, практически невозможно. Поэтому необходимо перейти к вычислению вероятности $P(x < 3)$. Она равна:

$$P(x = 0) + P(x = 1) + P(x = 2).$$

Откуда:

$$P(x \geq 3) = 1 - P(x < 3) = 1 - \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} =$$

$$= 1 - e^{-\lambda}\left(1 + 1{,}18 + \frac{1{,}8^2}{2}\right) = 1 - e^{-1{,}8} \cdot 4{,}42 =$$

$$= 1 - \frac{4{,}42}{e^{1{,}8}} = 1 - \frac{4{,}42}{2{,}718^{1{,}8}} = 0{,}27.$$

Характерной особенностью распределения Пуассона является то, что дисперсия распределения D равняется среднему значению λ. В том случае, когда требуется найти вероятность успешного исхода в схеме Бернулли, при количестве испытаний $n > 100$ и вероятности успеха $P \leq 0{,}1$, можно пользоваться формулой Пуассона, которая будет выглядеть так:

$$P(x) = \frac{(n \cdot P)^x e^{-np}}{x!},$$

где $n \cdot P$ — среднее значение, равное λ.

На этом основании закон Пуассона иногда называют законом редких явлений. Замену рекомендуется применять и при ручном вычислении вероятности биномиально распределенной случайной величины, так как формула Пуассона в вычислительном отношении проще формулы Бернулли (2.9).

В то время как дискретные случайные величины являются результатом подсчетов, результаты различных измерений представляются непрерывными случайными величинами. На горизонтальной оси наблюдений они изображаются не отдельными точками, как дискретные величины, а интервалами непрерывных точек.

Одним из главных законов распределения вероятностей непрерывной случайной величины, широко используемым в статистике, является нормальный закон распределения, часто называемый законом Гаусса. Основная особенность этого закона, выделяющая его среди других законов распределения, состоит в том, что к нему при определенных условиях приближаются многие другие законы.

Оказывается, что сумма большого числа независимых случайных величин, подчиненных каким-либо законам распределения, приближенно представляется нормальным законом и ошибка приближения тем меньше, чем больше суммируется этих случайных величин и среди них нет грубых выбросов.

Практика показывает, что такая закономерность характерна для многих явлений природы и общества. Поэтому распределения вероятностей, получаемые на основании статистического изучения этих явлений, чаще всего представляются нормальным законом.

Это одна из форм центральной предельной теоремы, играющей большую роль в теории вероятностей.

В связи с тем что для непрерывной случайной величины из определенного интервала, например $0 \leq x < 10$, нельзя указать

все точки, включенные в этот интервал, а следовательно, и соответствующие им вероятности, нормальное распределение представляется кривой, изображенной на рис. 2.2.

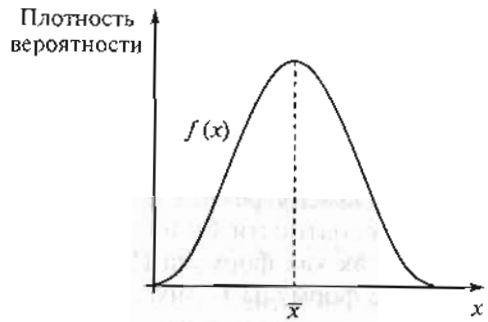

Рис. 2.2. Кривая нормального распределения

Эта кривая имеет колоколообразный характер, симметрична, с нормальным эксцессом. Это так называемая кривая плотности вероятности, нормального распределения $f(x)$, т. е. вероятности, приходящейся на единицу длины, когда последняя стремится к нулю.

Необходимость рассмотрения такого понятия вызвана тем, что точка на оси Ox не имеет геометрических размеров, следовательно, для нее нельзя указать конкретную вероятность. Такую вероятность ΔP можно указать только на отрезке Δx оси Ox. Если полученную вероятность отнести к длине Δx, т. е. положить $\frac{\Delta P}{\Delta x}$, и устремить Δx к нулю, пытаясь тем самым получить вероятность конкретной точки x, получим плотность вероятности $f(x)$.

Максимальная ордината кривой вероятности нормального закона соответствует точке оси x, определяемой средним арифметическим $\bar{X}$, медианой X_m и модой X_{md}.

По мере удаления от этой точки плотность распределения стремится к нулю. Площадь, заключенная под кривой, равна 1. Кривая описывается уравнением

$$f(x, \bar{X}, S) = \frac{1}{S\sqrt{2\pi}} e^{-\frac{(x-\bar{X})^2}{2S^2}}, \qquad (2.10)$$

где x — варианта;

S — стандартное отклонение;

π, e — константы, равные 3,14 и 2,718;

$\bar{X}$ — среднее арифметическое.

Изменение среднего арифметического выборки $\bar{X}$ при постоянстве стандартного отклонения S приводит к смещению кривой вдоль оси абсцисс. С увеличением S кривая становится более пологой, с уменьшением S — более острой.

Знание среднего арифметического $\bar{X}$ и стандартного отклонения S для нормально распределенной величины дает возможность определить пределы ее возможных значений. Практика показывает, что большинство нормально распределенных случайных величин лежат в интервале $[\bar{X} - 3S, \bar{X} + 3S]$. В связи с тем что для параметров генеральной совокупности приняты эквивалентные обозначения $\bar{X}$ — это μ (мю), а S — это σ (сигма), указанное правило получило название «правила трех сигма». На этом основании функцию (2.10) принято записывать так:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \qquad (2.11)$$

В связи с тем что площадь, охватываемая кривой $f(x, \mu, \sigma)$, представляет собой вероятность $P = 1$, любой участок этой площади, определяемый интервалом $x + \Delta x$ и кривой, представляет вероятность попадания случайной x в указанный интервал. Эту вероятность весьма часто приходится вычислять при решении различных задач статистики. В свою очередь, указанное вычисление сводится к определению части площади кривой, опирающейся на интервал $x + \Delta x$.

Оказалось, что вычисление указанной площади проще всего может быть осуществлено при помощи специальной функции Лапласа, для значений которой предварительно составляются таблицы. На этом основании вероятность того, что случайная величина x попадет в интервал (a, b), т. е. $P(a < x < в)$, через функцию Лапласа Φ^* определяют по следующему выражению:

$$P(a < x < b) = \Phi^*\left(\frac{b-\mu}{\sigma}\right) - \Phi^*\left(\frac{a-\mu}{\sigma}\right). \qquad (2.12)$$

Например, вероятность того, что случайная величина x будет принадлежать интервалу $(\mu - 2\sigma, \mu + 2\sigma)$, находится так:

$$\Phi^*\left(\frac{\mu + 2\sigma - \mu}{\sigma}\right) - \Phi^*\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right) = \Phi^*(2) - \Phi^*(-2).$$

Поскольку функция Лапласа нечетная $\Phi^*(-2) = -\Phi^*(2)$, согласно таблице ее значений [2], $\Phi^*(2) = 0{,}4772$. Таким образом, окончательно получаем $P(\mu - 2\sigma; \mu + 2\sigma) = 0{,}4772 + 0{,}4772 \approx 0{,}95$.

Аналогичным образом можно получить вероятность того, что случайная величина x будет принадлежать интервалу $(\mu - \sigma, \mu + \sigma)$ или интервалу $(\mu - 3\sigma, \mu = +3\sigma)$.

Для первого интервала $(\mu - \sigma, \mu + \sigma)$ имеем:

$$\Phi^*\left(\frac{\mu + \sigma - \mu}{\sigma}\right) - \Phi^*\left(\frac{\mu - \sigma - \mu}{\sigma}\right) = \Phi^*(1) - \Phi^*(-1) = 2\Phi^*(1) =$$

$$= 2 \cdot 0{,}3413 = 0{,}6826.$$

Для второго интервала $(\mu - 3\sigma, \mu = +3\sigma)$ получаем $2\Phi^*(3) = 0{,}997$.

Полученные вероятности 0,6826, 0,95, 0,997 означают, что примерно 68, 95 и 99 % наблюдений нормально распределенной случайной величины будут соответственно находиться в интервалах (–1, 1), (–2, 2), (–3, 3) стандартных отклонений σ.

Эти расчеты подтверждаются во многих практических случаях, когда случайная величина имеет нормальное распределение вероятностей. Когда же эта величина распределена не нормально, для определения процентных принадлежностей случайной, заданных тем или иным интервалом, можно использовать соотношение Чебышева.

Согласно его теореме для любого числа $k > 1$ по меньшей мере $\left(1 - \frac{1}{k^2}\right) \cdot 100$ % значений случайной попадут в k стандартных отклонений от средней.

Таким образом, при $k = 2$ по крайней мере 75 % всех значений случайной величины попадут в интервал двух стандартных отклонений от средней $\left(1 - \frac{1}{2^2} = -\frac{1}{4} = 0{,}75\right)$.

При $k = 3$ по формуле Чебышева получаем $1 - \frac{1}{9} = \frac{8}{9} = 0{,}89$, т. е. 89 % случайной попадут в интервал трех стандартных отклонений.

При $k = 4$ как минимум 93,7 % окажутся в пределах четырех стандартных отклонений от средней.

Поскольку многие аналитические методы статистики опираются на предположение, что случайная величина распределена нормально, на практике всегда следует убеждаться, что это так. Прикидочные проверки можно осуществить по статистикам средних, асимметрии и эксцесса. Другие методы проверки можно найти в [3, 4]. Фундаментальная проверка будет изложена в параграфе 3.3.

Величины $b - \mu$, $a - \mu$ (формула 2.12) представляют собой отклонения значений случайных a, $в$ от значения средней арифметической μ. Величины $\frac{b - \mu}{\sigma}$, $\frac{a - \mu}{\sigma}$ — относительные значения, выраженные в единицах среднего квадратического отклонения σ.

Обычно в статистике эти величины принято обозначать символом $Z = \frac{x - \mu}{\sigma}$, где x — значение изучаемой случайной величины.

Поэтому таблицы, согласно которым вычисляются вероятности по формуле (2.12), составлены для разных значений Z. Они называются стандартными, а распределение, выраженное в Z, стандартным с $\mu = 0, \sigma = 1$. Геометрически оно представляет доли площадей, заключенных между μ и Z, ограниченные кривой нормального распределения (рис. 2.3).

Тогда вероятность $P(a \le x \le b)$ равна сумме долей площадей, определяемых $Z_1 = \frac{b - \mu}{\sigma}$, $Z_2 = \frac{\mu - a}{\sigma}$.

В данном случае при $\mu = 100$, $b = 130$, $a = 55$, $\sigma = 15$ получаем

$$Z_1 = \frac{130 - 100}{15} = 2, \quad Z_2 = \frac{100 - 55}{15} = 3.$$

Согласно стандартным таблицам доли площади для $Z_1 = 0{,}4772$, для $Z_2 = 0{,}4986$. Поэтому

$$P(a \le x \le b) = 0{,}4772 + 0{,}4986 = 0{,}9758.$$

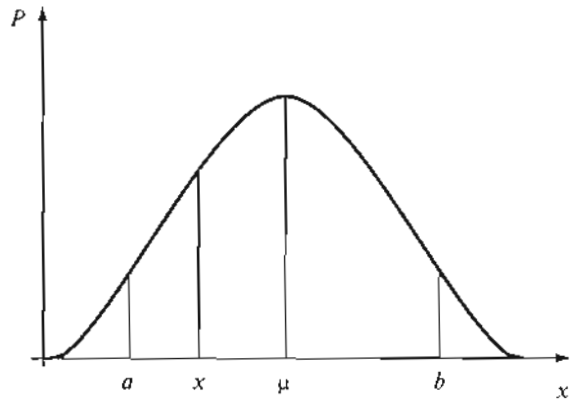

Рис. 2.3. Вычисление вероятности по значению Z

Таким образом, ценность нормального распределения вероятности изучаемой случайной величины x состоит в том, что при известном среднем значении этой случайной μ всегда можно сказать, что вероятность исхода x будет меньше или больше μ. Если же известно среднее квадратическое отклонение σ, то при заданном x можно установить вероятность того, что она принадлежит некоторому интервалу (a, b).

Используя стандартное нормальное распределение Z, вычислим, например, вероятность того, что случайная нормально распределенная величина x будет принадлежать интервалу $(-1, 1)$, т. е. находиться в пределах одного σ. Рисунок 2.4 демонстрирует этот случай.

Для того чтобы вычислить такую вероятность, необходимо определить вероятность $P(x \le 1)$. Согласно таблице стандартных значений эта вероятность равна 0,8413. Так как общая вероятность, определяемая кривой нормального распределения, равна 1, то вероятность $P(x > 1)$ равна $1 - 0{,}8413 = 0{,}1587$. Так как кривая симметрична, то вероятность $P(x < -1)$ тоже равна 0,1587. Поэтому вероятность попадания x в интервал $(-1, 1)$ равна $1 - 2 \cdot 0{,}1587 = 0{,}6826$.

Это то же значение, которое было получено раньше по формуле Лапласа для интервала $(\mu - \sigma, \mu + \sigma)$.

В заключение отметим, что при решении различных задач используют и функцию накопленных частот (кумуляту) нормального распределения. Через функцию Лапласа она выражается так:

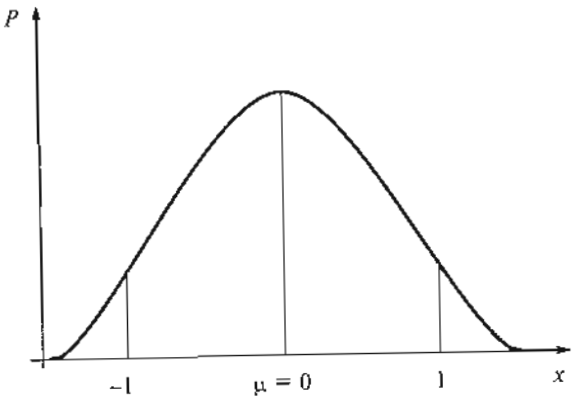

Рис. 2.4. К вычислению $P(-1 \le x \le 1)$

$$F(x, \bar{X}, \sigma) = \Phi^{*}\left(\frac{x - \mu}{\sigma}\right). \qquad (2.13)$$

2.3. Подготовка выборки и выборочные распределения

Для того чтобы на основе исследования выборочной совокупности наблюдений можно было сделать правдоподобные статистические заключения о свойствах элементов генеральной совокупности, выборка должна быть специальным образом подготовлена. Говорят, она должна быть **репрезентативной**, т. е. представительной.

В свою очередь, репрезентативность выборки в процессе ее подготовки обеспечивается выполнением ряда условий. Прежде всего количество элементов выборки n должно быть достаточным. Однако именно это требование трудновыполнимо.

В учебниках по математической статистике, например [6], приводится формула, согласно которой количество элементов выборки может быть найдено, если известна дисперсия генеральной совокупности. Однако часто дисперсия требует оценки на основе выборки.

Таким образом, получается замкнутый круг: объем выборки может быть найден, если известна дисперсия, а дисперсия найдена на основании выборки.

Поэтому при определении объема выборки на практике придерживаются следующего правила: чем больше n, тем достовернее статистические заключения о свойствах элементов генеральной совокупности. Это правило следует из закона больших чисел.

Обычно большой считается выборка, содержащая 100 и более элементов. Ряд статистиков полагают выборку достаточной, если она содержит 30—50 элементов. Выборка, содержащая меньше 30 элементов, является маленькой.

Второе правило, которому нужно следовать при формировании представительной выборки, — элементы выборки должны быть частью элементов генеральной совокупности, относительно характеристик которой требуется сделать выводы на основе изучения выборки.

Например, фирма НОКИА, изготавливающая мобильные телефоны разных уровней сложности, выбирает для определения срока службы 100 телефонов простейшей конструкции и делает заключение о среднем сроке службы телефонов разных конструкций. Такое использование выборки ошибочно. Точное заключение может быть сделано только о телефонах простейшей конструкции.

Наконец, третье правило формирования репрезентативной выборки — она должна быть случайной, часто говорят, вероятностной.

Случайность означает то, что при составлении выборки каждый ее элемент должен иметь равную с другими элементами возможность попасть в выборку. Такая возможность достигается разными способами. Простейший из них — случайный выбор. Он неоднократно демонстрировался на экранах телевизоров в розыгрышах спорт-лотереи, капитал шоу «Поле чудес» и др.

Предположим, некоторый завод производит электродвигатели в объеме 500 штук ежемесячно. На контроль исправности отправляют 50 двигателей. Требуется, чтобы выборка $n = 50$ была случайной.

Пронумеруем все двигатели числами от 1 до 500 и запишем эти числа на картонных жетонах. Поместим жетоны в некоторую емкость, тщательно их перемешаем и последовательно вынем из

емкости 50 жетонов. Номера этих жетонов укажут двигатели, которые необходимо включить в выборку.

В настоящее время программное обеспечение всех компьютеров включает генераторы случайных чисел, которые можно использовать для формирования случайных выборок.

Другой способ получения случайной выборки — систематический отбор, когда из генеральной совокупности выбирается каждый j-й элемент. Например, из 500 студентов в выборку 50 человек назначается каждый 10-й студент.

В практической деятельности для составления случайной выборки весьма часто применяют групповой отбор, когда генеральная совокупность разбивается на равновеликие группы и из каждой группы систематически выбирается заданное количество элементов. Есть и другие способы получения репрезентативных выборок. Тем не менее при всех методах формирования выборки, поскольку в статистическом заключении статистик опирается на выборку, это вносит вероятность ошибки в это заключение, принято говорить, ошибку выборки.

Рассмотрим следующую процедуру. Из некоторой совокупности, например 5000 элементов последовательно берется 100 случайных выборок объемом по 50 элементов каждая. Для каждой выборки подсчитываются статистики: среднее арифметическое, стандартное отклонение, медиана, мода и т. д.

В общем случае значения соответствующих статистик выборок будут различаться между собой и на множестве 100 выборок их можно рассматривать как случайные величины.

Для этих случайных величин правомерно рассматривать свои статистики, например говорить о среднем арифметическом среднего выборок, о стандартном отклонении этого среднего, о медиане и моде среднего. Также можно рассматривать среднее арифметическое стандартных отклонений выборок, стандартное отклонение этого среднего и т. д.

Правомерно говорить и о распределениях частот статистик выборок. Например, можно рассматривать распределение частот средних арифметических выборок и характеризовать его средним арифметическим средних и стандартным отклонением средних.

В отличие от распределения отдельных наблюдений такие распределения частот принято называть выборочными. Характерной особенностью этих распределений является то, что даже в том случае, когда распределение исходной совокупности на-

блюдений, в данном случае 5000 элементов, не соответствует нормальному закону, выборочные распределения при увеличении объема выборок приближаются к нормальному распределению частот, а следовательно, и вероятностей.

На рис. 2.5 показаны кривые распределения вероятного распределения средней арифметической, изменяющие свою форму при увеличении объемов выборок. При этом исходным является равномерное распределение.

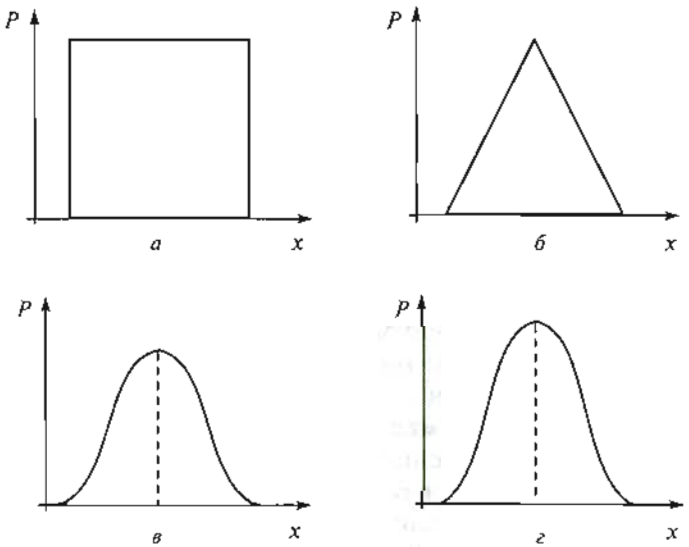

Рис. 2.5. Совокупность наблюдений с равномерным распределением:
а — выборочные распределения среднего арифметического $\bar{X}$; *б* — при $n = 2$; *в* — при $n = 5$; *г* — при $n = 30$

При объеме выборки $n = 30$ выборочное распределение среднего арифметического $\bar{X}$ практически нормальное. Строгое доказательство этого положения приводится в учебниках по теории вероятностей и формулируется в виде центральной предельной теоремы. Практически оно дает возможность пользоваться нормальным законом распределения вероятностей в процессе различных статистических заключений, опираясь при этом на выборочные распределения.

Безусловно, задача формирования множества выборок для составления выборочного распределения, например среднего

арифметического выборки, является достаточно трудоемкой. Однако имеются обоснованные формулы для оценки характеристик выборочных распределений. В связи с этим подготовка многих случайных выборок необязательна.

Практически подтверждено, что выборочное среднее среднего арифметического ($\mu_{\bar{x}}$) многих выборок приближается к среднему арифметическому генеральной совокупности. Иными словами, справедливо равенство

$$\mu_{\bar{x}} = \mu. \tag{2.14}$$

Стандартное отклонение $\sigma_{\bar{x}}$ выборочного среднего связано со стандартным отклонением генеральной совокупности таким отношением:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{2.15}$$

Из этой формулы следует, что $\sigma_{\bar{x}}$ всегда меньше σ. При этом, когда рассматривается конечная генеральная совокупность, в указанную формулу вносится поправка на число элементов N этой совокупности. Вследствие этого формула (2.15) приобретет такой вид:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}. \tag{2.16}$$

Принято считать, что поправку на число элементов генеральной совокупности вносят тогда, когда объем выборки n составляет более 5 % числа элементов этой совокупности. Например, если $N = 100$, а $n = 48$, вычисление σ необходимо вести по формуле (2.16).

Вычисление характеристик выборочных распределений $\mu_{\bar{x}}$ и $\sigma_{\bar{x}}$ по параметрам генеральной совокупности не всегда осуществимо, так как во многих случаях эти параметры неизвестны и требуют оценки. В том случае, когда подготовлена случайная выборка, лучшая оценка среднего арифметического генеральной совокупности — это $\mu = \bar{X}$.

Оценка стандартного отклонения генеральной совокупности по статистике выборки и числу ее элементов n определяется по

выражению $\sigma = S\sqrt{\frac{n}{n-1}}$. Подставляя это выражение в формулу (2.15) для оценки стандартного отклонения выборочного распределения среднего арифметического, получим:

$$S_{\bar{x}} = \frac{S\sqrt{n/(n-1)}}{\sqrt{n}} = \frac{S}{\sqrt{n-1}}. \qquad (2.17)$$

Таким образом, оценка среднего квадратического отклонения среднего арифметического выборочного может быть вычислена по четырем формулам: (1.15), (1.16), если известно стандартное отклонение генеральной совокупности σ, n, и (2.17), (2.18), если известно стандартное отклонение выборки S объема n. При этом формула (2.18) имеет вид:

$$S_{\bar{x}} = \frac{S}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}} \qquad (2.18)$$

и употребляется в том случае, когда величина n составляет больше 5 % от числа элементов генеральной совокупности N.

Подобно всякому среднему квадратическому отклонению, среднее квадратическое статистик выборки характеризует степень изменчивости выборочного распределения. Оно представляет как бы ошибки выборки при оценке параметров генеральной совокупности. Поэтому среднее квадратическое отклонение или стандартное отклонение средней арифметической выборок называют стандартной ошибкой средней арифметической при оценке средней генеральной совокупности μ на основании выборочной средней $\mu_{\bar{x}}$.

Аналогично, рассматривая распределение медиан, его можно было бы характеризовать средним арифметическим медиан и стандартной ошибкой медиан. Выборочное распределение средних квадратических отклонений можно было бы характеризовать их средним значением и стандартной ошибкой среднего квадратического отклонения.

На применении стандартных ошибок средних $\sigma_{\bar{x}}$, $S_{\bar{x}}$ построен практически весь статистический анализ. Они широко используются при вычислении интервальных оценок параметров генеральной совокупности, проверки гипотез и др.

2.4. Компьютерные технологии формирования законов распределения и случайных выборок

Компьютерные технологии формирования законов распределения и подготовки случайных выборок в пакете **Анализ данных** реализуется инициацией двух пунктов меню пакета — **Генерация случайных чисел** и **Выборка**, а также вычислением значений ряда статистических функций.

Пункт **Генерация случайных чисел** позволяет сформировать массив случайных чисел, распределенных по указанному закону. Для этого после вывода на экран меню пакета **Анализ данных** необходимо выделить пункт меню **Генерация случайных чисел** и щелкнуть мышью кнопку **ОК**. Далее в появившемся диалоговом окне необходимо выделить тип распределения и снова щелкнуть кнопку **ОК**. После этого необходимо заполнить следующие поля:

1) *число переменных* — указывается число столбцов книги Excel, в которых будут хранить сгенерированные случайные числа;

2) *число случайных чисел* — указывается то их число, которое необходимо сгенерировать;

3) *параметры* — указываются значения, определяемые законом распределения; например, для нормального закона указываются среднее μ и стандартное отклонение σ либо остается 0 или 1 по умолчанию; для биномиального закона указываются вероятность успеха P и число испытаний n;

4) *случайное рассеивание* — указывается число раскачки генератора случайных чисел, например 20, 50, 100 и т. д.

Число переменных, т. е. число столбцов, где будут храниться сгенерированные числа, увеличивает количество этих чисел в целое число раз. Например, если число случайных чисел 30, а число столбцов 3, то всего будет сгенерировано 90 чисел.

Параметр *случайное рассеивание* необходим для того, чтобы при повторной генерации получить один и тот же набор случайных чисел.

Для построения графиков распределения вероятностей на базе сгенерированных наборов случайных чисел необходимо прежде всего упорядочить наборы этих чисел по возрастанию.

Далее при помощи функции, определяющей вероятности для каждого элемента полученного массива, вычислить соответствующую вероятность и получить таким образом набор вероятностей. После этого при помощи **Мастера диаграмм** построить график распределения вероятностей. Например, требуется построить график нормального распределения предварительно полученного согласно пункту **Генерация случайных чисел** набора 30 случайных чисел, размещенных в ячейках A1:A30.

Средствами Excel упорядочиваем эти числа. Затем при помощи функции **Нормрасп**, имеющейся в составе статистических функций Excel, последовательно по формуле (2.10) вычисляем плотность вероятности диапазона случайных A1:A30.

Далее при помощи **Мастера диаграмм** строим график распределения. Интегральная функция **Нормрасп**($(x, \bar{X}, S, 1)$), в которой x — случайное из нормального распределения; $\bar{X}$ — среднее арифметическое распределения; S — стандартное отклонение; 1 или 0 — признак того, что требуется строить интегральную функцию вероятности, к которой стремится кумулята частот, или функцию плотности вероятности.

Функция **Нормрасп** ($(x, \bar{X}, S, 1)$) может быть использована и для определения вероятности того, что случайная величина попадает в заданный интервал (a, b), рассчитываемый по формуле (2.12). Проиллюстрируем это на примере, взятом из [5].

Для продажи мужских курток в некотором городе торговая фирма провела выборочное обследование роста мужчин в возрасте от 18 до 65 лет. В результате было установлено, что закон распределения нормальный, средний рост мужчин $\bar{X} = 176$ см, а стандартное отклонение $S = 6$ см.

Необходимо определить, какой процент общего числа курток должны составлять куртки 5-го размера (182—185 см).

Формула для решения задачи в соответствии с (2.12) имеет такой вид:

$$\text{Нормрасп}(186,176,6,1) - \text{Нормрасп}(182,176,6,1) =$$

$$= 0{,}95221 - 0{,}84134 = 0{,}11086 \approx 11\ \%.$$

Таким образом, куртки 5-го роста должны составлять приблизительно 11 % общего числа курток.

В составе статистических функций Excel часто используется функция **Нормобр**(P, $\bar{X}$, S), вычисляющая значения случайной величины (квантиль) по заданной ее вероятности P.

В частности, квантиль уровня 1/2 есть медиана, квантили 1/4, 3/4 — квартили, квантили 1/10, 2/10, ..., 9/10 — децили. Для работы с этой функцией необходимо задать: вероятность $P(x)$, среднее арифметическое выборки $\bar{X}$ и стандартное отклонение S.

Функция **Биномрасп**(*число успехов x*, *число испытаний n*, *вероятность успеха P* (*x*), *интегральная*) предназначена для вычисления вероятности x успешных исходов по формуле (2.9), характеризующей биномиальное распределение. Если аргумент интегральная = 1, рассчитывается вероятность равно x успешных исходов.

Вычислив вероятности последовательности успешных исходов $x = 0, 1, 2, \ldots$ при помощи функции **Биномрасп**, можно получить биномиальный закон распределения, характеризующий ситуацию с заданным значением исхода $P(x)$.

Для вычисления вероятности распределения Пуассона в пакете **Анализ данных** предусмотрена функция **Пуассон**(*число реализаций x*, *среднее* $\bar{X}$, *интегральная*). В том случае, когда аргумент интегральная = 1, вычисляется вероятность того, что в заданный интервал времени произойдет не больше чем x реализаций; когда аргумент интегральная равен 0, вычисляется вероятность равно x реализаций.

Использование функций стандартное. Требуемая функция выделяется в списке статистических функций, и нажимается кнопка **ОК**. Далее в открывшемся окне заполняются поля — заносятся значения аргументов функций и снова нажимается кнопка **ОК**. В результате на экране появляется вычисленное значение соответствующей функции.

Пункт **Выборка** пакета **Анализ данных** позволяет сформировать выборку из набора данных, определяющих генеральную совокупность. Для этого необходимо выделить данный пункт (щелкнуть мышью), а затем нажать кнопку **ОК**.

В открывшемся окне в поле *входной интервал* следует указать диапазон данных генеральной совокупности, например A1:A300, инициировать пункт *случайный*, в поле *число выборок* указать число элементов выборки, например 30, и обозначить адрес вывода: выходной интервал или новый лист. После заполнения указанных полей необходимо нажать кнопку **ОК**.

Контрольные вопросы

1. Можно ли исключить риск сделать ошибку в процессе статистических заключений о свойствах генеральной совокупности наблюдений на основе статистической выборки?
2. Как можно уменьшить риск?
3. Что такое вероятность случайного события?
4. В каких пределах вычисляется вероятность?
5. Какие способы вычисления вероятности событий вам известны?
6. Объясните классический подход к вычислению вероятности и приведите пример.
7. Объясните частотный подход к вычислению вероятности. В чем различие между частотным и классическим подходами?
8. В чем смысл закона больших чисел и как он проявляется по отношению к увеличению частоты события?
9. Как формулируется правило сложения вероятностей для несовместимых и совместимых событий?
10. Как формулируется правило умножения вероятностей для зависимых и независимых событий?
11. Что такое субъективная вероятность?
12. Что понимается под законом распределения вероятности?
13. Можно ли оценить риски в принятии решений, используя значения вероятности?
14. Какую ситуацию отражает распределение: 1) биноминальное; 2) Пуассона?
15. Какие особенности отличают нормальный закон распределения вероятности от ряда других законов распределения?
16. Как изображается нормальный закон распределения?
17. Что такое плотность вероятности?
18. Какое уравнение описывает кривую плотности вероятности?

19. Как вычислить вероятность того, что случайная величина x принадлежит интервалу (a, b)?
20. Какие требования предъявляются к подготовке выборки наблюдений?
21. Что означает термин «случайная (вероятностная) выборка»?
22. Что представляют собой выборочные распределения? Как их можно получить?
23. Какова основная особенность выборочных распределений?
24. По каким формулам можно вычислить выборочное среднее арифметическое и стандартное отклонение выборочного среднего, если известны параметры генеральной совокупности и если они не известны?
25. Какие пункты меню пакета **Анализ данных** следует использовать для построения законов распределения случайных величин и подготовки случайных выборок?
26. Какие возможности предоставляет использование статистических функций **Нормрасп**, **Нормобр**, **Биномрасп**, **Пуассон**?

Задачи

1. Производится 100 опытов. Из них 60 удачных. Чему равны частоты удач и неудач? Запишите формулу вычисления относительной частоты и определите эту частоту.
2. В коробке 6 шаров: 2 белых и 4 черных. Определите вероятность вынимания шара любого цвета, белого шара, черного шара.
3. При бросании шестигранной игральной кости определите вероятность выпадения 1 очка, 6 очков.
4. Определите вероятность вытягивания туза или короля из колоды 52 карт.
5. Выполняется три последовательных подбрасывания монеты. Найдите вероятность выпадения орла, решки.
6. В состав партии входят 10 приборов, один из которых неисправный. Укажите вероятность обнаружения неисправ-

ного прибора при проверке двух приборов, если первый оказался неисправным.

7. В ящике 7 шаров: 3 белых и 4 черных. Определите вероятность вынимания подряд двух белых шаров.
8. Производится 10 опытов. Вероятность успешного исхода $P = 1/2$. Найдите вероятность появления пяти успешных исходов.
9. Вероятность успешного исхода $P = 0{,}05$. Проводится 200 опытов. Найдите вероятность 100 успешных исходов.
10. Для среднего арифметического $\bar{X} = 1$ и стандартного отклонения $\sigma = 1$. При помощи функции **Нормраспр** найдите вероятность попадания случайной величины x в интервалы $(\bar{X} - \sigma, \bar{X} + \sigma)$, $(\bar{X} - 2\sigma, \bar{X} + 2\sigma)$, $(\bar{X} - 3\sigma, \bar{X} + 3\sigma)$.
11. Найдити квантиль вероятности уровня 1/3.
12. Сгенерируйте набор из 30 нормально распределенных случайных чисел с $\bar{X} = 0, \sigma = 1$.
13. Сгенерируйте случайную выборку $n = 30$ из совокупности $N = 100$.

Глава 3
ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ И ПРОВЕРКА ГИПОТЕЗ

3.1. Точечные и интервальные оценки параметров

Числовые характеристики, полученные на основании выборки, называемые статистиками, также принято называть точечными оценками параметров генеральной совокупности. Объясняется это тем, что эти характеристики, являясь приближенными значениями параметров, как бы оценивают параметры, находясь слева или справа по отношению к их значениям.

Таким образом, точечные оценки — это единичные значения, лучше всего описывающие параметры генеральной совокупности. Для того чтобы оценки были более объективными, т. е. более точно оценивали параметры, они должны удовлетворять ряду требований. Эти требования следующие.

Оценки должны быть состоятельными, несмещенными и эффективными (результативными).

Оценка параметра является состоятельной, если по мере роста числа наблюдений она стремится к действительному значению этого параметра. Так, при $n \to \infty$ среднее арифметическое выборки $\bar{X}$ стремится к среднему арифметическому генеральной совокупности, т. е. значение $\bar{X} \to \mu$. Дисперсия D и стандартное отклонение S выборки стремятся соответственно к σ^2 и σ.

Оценка является несмещенной, если не содержит систематических отклонений. Другими словами, если при наличии большого числа оценок, полученных на основании многих выборок из генеральной совокупности, одинаковое число значений оце-

нок лежит слева и справа от значения параметра, оценка считается несмещенной. Оценка дисперсии S^2 является смещенной, так как ее среднее $S^2 = \sigma^2 - \frac{\sigma^2}{n}$ смещено на величину $\frac{\sigma^2}{n}$.

Таким образом, показатель смещенности $\frac{\sigma^2}{n}$ при увеличении n уменьшается и, наоборот, при уменьшении n возрастает. Поэтому при малых значениях n дисперсию D вычисляют по выражению $D = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$.

Оценка называется эффективной, если среди прочих оценок того же параметра она имеет наименьшую дисперсию, т. е. разброс.

Статистикой, которая удовлетворяет всем перечисленным требованиям, является среднее арифметическое выборки $\bar{X}$. Поэтому оно служит наилучшей оценкой параметра генеральной совокупности μ. Однако не каждая статистика выборки является наилучшей оценкой соответствующего параметра генеральной совокупности. Например, оценкой стандартного отклонения σ генеральной совокупности служит величина $S\sqrt{\frac{n}{n-1}}$, где $\sqrt{\frac{n}{n-1}}$ — поправочный коэффициент, устраняющий смещенность оценки.

Точечные оценки, используемые в статистике, — это оценки среднего, генеральной совокупности, стандартного отклонения $\sigma = S\sqrt{\frac{n}{n-1}}$, разности средних двух генеральных совокупностей $\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2$, суммы всех элементов генеральной совокупности $N\bar{X}$, долей и др.

Оценки определяются довольно просто, однако имеют существенный недостаток: никогда нельзя себе представить, насколько эти оценки точны. Поэтому в статистике широко используются *интервальные оценки*.

Эти оценки представляют собой длины интервалов, которые с наперед заданной вероятностью содержат значения параметров генеральной совокупности. Например, диапазон значений, который с вероятностью 0,95 может содержать среднее число генеральной совокупности, называется интервальной оценкой средней μ.

Такой интервал принято называть доверительным по той причине, что уровень (степень) доверия к его длине определен величиной вероятности, в данном случае ее значением 0,95. В свою очередь, это означает, что из 100 интервалов заданной длины 95 из них будут содержать среднее арифметическое генеральный совокупности μ.

Различие между точечной и интервальной оценкой продемонстрируем на примере. Некоторая фирма выпускает жидкокристаллические телевизионные экраны. Если, основываясь на исследовании выборки заданного объема n, средний срок службы экрана оценивают в 100 000 ч, то при этом используют точечную оценку. Если же с вероятностью 0,95 устанавливают, что средний срок службы экрана 80 000—120 000 ч, это заключение есть результат интервальной оценки.

Интервальные оценки вычисляют в предположении, что случайные данные имеют нормальный закон распределения. При этом для определения доверительного интервала принятого уровня доверия средней арифметической генеральной совокупности используют либо стандартное отклонение выборочной средней $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}$, либо стандартное отклонение генеральной совокупности $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Объясняется это тем, что средние арифметические ряда выборок, взятые из рассматриваемой совокупности наблюдений, имеют тенденцию к группировке вокруг значения μ, а $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, определяет колебания множества этих средних относительно μ. Поэтому в общем случае интервал колебаний $(\bar{X} - S_{\bar{x}}, \bar{X} + S_{\bar{x}})$ средних арифметических $\bar{X}$ либо $(\mu - \sigma_{\bar{x}}, \mu + \sigma_{\bar{x}})$ со средней μ и будет интервальной оценкой средней арифметической генеральной совокупности.

Теперь осталось определить уровень доверия к этому интервалу, т. е. установить вероятность, которая будет гарантировать, что из всех интервалов определенное их количество будет содержать среднюю арифметическую μ. Чаще всего на практике в ка-

честве указанной величины принимают значение 0,95, довольствуясь тем, что только пять интервалов из 100, т. е. 5 их процентов, не будут содержать μ.

Вместе с тем иногда могут потребоваться и более высокие уровни доверия: 0,97 и 0,99.

Как же влияет уровень доверия на величину интервала?

Безусловно, чем выше этот уровень, т. е. больше гарантия, тем шире должен быть интервал и тем менее точна интервальная оценка. Вместе с тем оставить тот же интервал при увеличении уровня доверия можно путем увеличения объема выборки n. Если используется стандартная ошибка выборки $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}$, то, увеличивая объем выборки, мы уменьшаем $S_{\bar{x}}$ и таким образом способствуем уменьшению интервала $(\bar{X} - S_{\bar{x}}, \bar{X} + S_{\bar{x}})$ либо $(\mu - \sigma_{\bar{x}}, \mu + \sigma_{\bar{x}})$.

Конкретное определение длины доверительного интервала относительно средней арифметической генеральной совокупности μ опирается на следующие рассуждения.

Практически установлено, что для нормального распределения случайной величины в среднем 68 %, 90, 95 и 99 % ее значений лежат соответственно в пределах:

$(\mu - \sigma_{\bar{x}}, \mu + \sigma_{\bar{x}})$, $(\mu - 1{,}64\sigma_{\bar{x}}, \mu + 1{,}64\sigma_{\bar{x}})$,

$(\mu - 1{,}96\sigma_{\bar{x}}, \mu + 1{,}96\sigma_{\bar{x}})$, $(\mu - 2{,}58\sigma_{\bar{x}}, \mu + 2{,}58\sigma_{\bar{x}})$ (рис. 3.1).

С другой стороны, поскольку площадь кривой, опирающаяся на интервал, есть вероятность того, что случайная величина попадает в этот интервал, то 68—99 % — суть вероятности принадлежности случайных соответствующим интервалам.

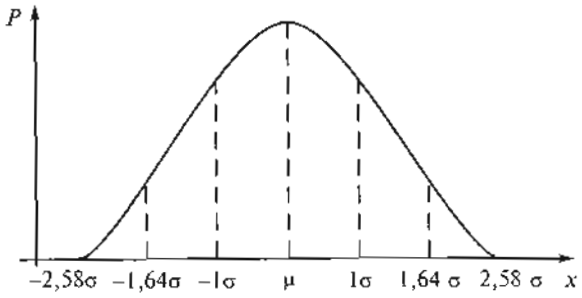

Рис. 3.1. Интервал размещения долей случайной для нормального распределения

Таким образом, если устанавливается уровень доверия, например, 0,95, то ему соответствует интервал $(\mu - 1{,}96\sigma, \mu + 1{,}96\sigma)$. И наоборот, если выбран интервал $(\mu - 1{,}96\sigma, \mu + 1{,}96\sigma)$, то с вероятностью 0,95 случайные величины будут принадлежать этому интервалу.

В связи с тем что в данном случае речь идет об интервальной оценке среднего арифметического генеральной совокупности, колебания которого как случайной величины более точно отображаются выборочным стандартным отклонением $S_{\bar{x}}$, интервальная оценка на основании выборки объема n для μ определяется так: $(\bar{X} - 1{,}96S_{\bar{x}}, \bar{X} + 1{,}96S_{\bar{x}})$.

Рассмотрим пример. В некоторой фирме работает 626 человек. Составлена случайная выборка из 50 рабочих, на основании которой определены статистики: средний недельный заработок $\bar{X} = 3300$ руб. и стандартное его отклонение $S = 250$ руб. Требуется определить 95%-ный доверительный интервал недельного заработка всех рабочих фирмы.

Поскольку указанный доверительный интервал определяется по формуле $(\bar{X} - 1{,}96S_{\bar{x}}, \bar{X} + 1{,}96S_{\bar{x}})$, требуется вычислить $S_{\bar{x}}$. Как известно, в зависимости от объема выборки $S_{\bar{x}}$ вычисляется по формуле $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}}$, если $n > 5\,\%N$ (формулы (2.17), (2.18)).

В нашем случае объем выборки $n = 50 > 5$ % числа 626, поэтому

$$S_{\bar{x}} = \frac{250}{49}\sqrt{\frac{626-50}{625}} = \frac{250}{49}\cdot\frac{24}{25} = \frac{240}{49} = 4{,}898.$$

Теперь можно определить левую и правую границы интервала. Левая граница $\bar{X} - 1{,}96S_{\bar{x}} = 3300 - 1{,}96 \cdot 4{,}898 = 3300 - 9{,}7 = 3290{,}3$.

Правая граница равна 3300 + 9,7 = 3309,7.

Таким образом, средний недельный заработок рабочих фирмы с вероятностью 0,95 колеблется в пределах от 3290,3 до 3309,7 руб.

Весьма часто управляющему фирмой необходимо знать о средних расходах, которые несет фирма за неделю по заработной плате. В этом случае можно воспользоваться точечной оценкой $\bar{X} = \mu$, и в данном случае это составит $N \cdot \bar{X} = 626 \cdot 3300 = 2065800$ руб.

Когда же нужно определить 95%-ный доверительный интервал колебаний общей зарплаты, необходимо воспользоваться интервальной оценкой $(N(\bar{X} - 1{,}96S_{\bar{x}}),\ N(\bar{X} + 1{,}96S_{\bar{x}}))$. Другими словами, найденные ранее границы интервала следует умножить на число служащих фирмы.

В общем случае формула определения доверительного интервала любой статистики генеральной совокупности такая: точечная оценка $\pm\sigma \cdot S_{\bar{S}i}$, где $S_{\bar{S}i}$ — стандартное отклонение соответствующей статистики. Обычно эта формула записывается так, что вместо σ используется стандартизованная величина $Z = \frac{\mu - \bar{X}}{\sigma}$, в единицах которых представляется натуральная случайная величина. Поэтому окончательно выражение для определения интервальной оценки будут иметь такой вид:

$$\text{Точечная оценка} \pm Z \cdot S_{\bar{S}i}. \qquad (3.1)$$

При этом вероятностям 0,68, 0,9, 0,95, 0,99 соответствуют такие значения Z: 1; 1,64; 1,96; 2,58. Поэтому, рассматривая тот или иной интервал доверия, мы всегда выбираем соответствующее численное значение Z.

Величину $Z \cdot S_{\bar{S}i}$ часто называют предельной ошибкой выборки и обозначают $\Delta_{\bar{x}}$ по той причине, что $S_{\bar{S}i}$ — средняя ошибка выборки соответствующей статистики, а Z — предельный коэффициент, определяемый уровнем доверия.

Ранее было рассмотрено построение 95%-ного доверительного интервала для среднего арифметического генеральной совокупности μ на основании средней выборки $\bar{X}$ и стандартного выборочного отклонения $S_{\bar{x}}$.

Теперь рассмотрим задачу построения доверительного интервала для разности средних арифметических двух выборок $\bar{X}_1$, $\bar{X}_2$.

Для ее решения необходимо составить две случайные выборки, вычислить статистики $\bar{X}_1$ и $\bar{X}_2$, найти выборочные стандартные отклонения $S_{\bar{x}_1}$ и $S_{\bar{x}_2}$. Затем определить точечную оценку разности средних арифметических $\Delta_{\bar{x}_1 - \bar{x}_2} = \bar{X}_1 - \bar{X}_2$ и составить формулу $\Delta_{\bar{x}_1 - \bar{x}_2} \pm Z \cdot S_{\bar{S}i}$.

Вычисление стандартного выборочного отклонения разности $S_{\bar{S}i}$ осуществляется по выражению $S_{\bar{S}i} = \sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}}$. Поэтому

доверительный интервал разности $\Delta_{\bar{x}_1 - \bar{x}_2}$ двух средних будет иметь такой вид:

$$\left(\Delta_{\bar{x}_1 - \bar{x}_2} - Z\sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}},\ \Delta_{\bar{x}_1 - \bar{x}_2} + Z\sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}}\right). \qquad (3.2)$$

Рассмотрим две фирмы. В одной из них работает 760 рабочих, в другой 626. По данным случайной выборки из 70 рабочих первой фирмы средняя зарплата $\bar{X}_1$ за неделю составила 3800 руб. при стандартном выборочном отклонении $S_{\bar{x}_1}$, равном 5,788 руб. По данным случайной выборки из 50 рабочих второй фирмы средняя зарплата $\bar{X}_2$ за неделю составила 3300 руб. при стандартном выборочном отклонении $S_{\bar{x}_2} = 4{,}898$ руб.

Требуется оценить наиболее вероятную разницу в недельной заработной плате рабочих двух фирм.

Точечная оценка разности средних зарплат $\Delta_{\bar{x}_1 - \bar{x}_2} = 3800 - 3300 = 500$. Стандартное выборочное отклонение разности:

$$S_{\bar{S}i} = \sqrt{(5{,}788)^2 + (4{,}898)^2} = \sqrt{33{,}5 + 23{,}99} = \sqrt{57{,}49} = 7{,}58 \text{ руб.}$$

Поэтому доверительный интервал для разности средней $\bar{X}_1 - \bar{X}_2$ будет таким: $(500 - Z \cdot 7{,}58; 500 + Z \cdot 7{,}58)$. Конкретно 95%-ный доверительный интервал имеет следующие границы: $500 - 1{,}96 \cdot 7{,}58 = 500 - 14{,}9 = 485{,}1$ и $500 + 14{,}9 = 514{,}9$.

Таким образом, разность средних двух выборок с вероятностью 0,95 заключена в пределах от 485,1 до 514,9 руб.

Рассмотрим следующий пример. Из 10 студентов университета, входящих в случайную выборку, четверо оказались курящими. Задача состоит в том, чтобы на основании этих данных дать правдоподобную интервальную оценку доли (части) курящих студентов всего университета.

Порядок решения этой задачи такой. Доля студентов выборки, курящих сигареты, составляет $P = \frac{4}{10} = 0{,}4$, или 4 %. Примем эту величину в качестве наилучшей точечной оценки соответствующей доли генеральной совокупности, т. е. всех студентов университета. Обозначив долю генеральной совокупности символом π, получим $\pi = P = 0{,}4$. Тем самым мы говорим, что, например, из 10 000 студентов 4000 являются курящими.

Теперь приступим к определению интервальной оценки доли генеральной совокупности. Раньше для определения интервальных оценок использовалась кривая нормального распределения вероятности в предположении, что рассматривались непрерывные величины, распределения которых согласно центральной предельной теореме аппроксимируются (приближаются) нормальным законам Гаусса.

В данном случае рассматриваются доли, которые определяются подсчетом их количеств (4 студента из 10), т. е. на самом деле имеем дело с дискретной случайной величиной. Поэтому по всем правилам для определения интервальной оценки доли следовало бы использовать закон распределения этой случайной величины. В частности, в рассматриваемом примере нужно использовать биномиальное распределение, определяемое коэффициентами разложения бинома $(a + b)$, где $a = 0{,}4$, $b = 0{,}6$, поскольку имеется два исхода: курит—не курит. Окажется ли это распределение симметричным, заранее неизвестно. В связи с этим пользоваться им для определения интервальной оценки достаточно сложно.

На практике найден следующий выход. Вследствие того, что с увеличением объема выборки характеристики биномиального распределения приближаются к характеристикам нормального, для поиска интервальной оценки доли генеральной совокупности пользуются нормальным распределением вероятности. При этом соблюдают следующие условия: объем выборки n должен быть не меньше 50, т. е. $n \geq 50$ и $n \cdot a \geq 5$, $n \cdot b \geq 5$.

Таким образом, в рассматриваемом примере число случайно отобранных студентов должно быть не меньше 50, что даст $n \cdot a = n \cdot p = 50 \cdot 0{,}4 = 20 > 5$ и $n \cdot b = 50 \cdot 0{,}6 = 30 > 5$.

Теперь, опираясь на основную формулу (3.1) построения доверительного интервала — точечная оценка $\pm Z \cdot S_{\bar{s}}$, для определения интервальной оценки доли генеральной совокупности можно записать:

$$\pi \pm Z \cdot S_{\bar{s}}. \qquad (3.3)$$

Конкретные значения границ интервалов можно получить, зная значения среднего квадратичного выборочного отклонения доли. Оно определяется по формуле

$$S_{\bar{s}} = \sqrt{\frac{P(1-P)}{n}},$$

если генеральная совокупность бесконечно большая, или по формуле

$$S_{\bar{s}} = \sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}},$$

если объем выборки составляет более 5 % N.

На основании изложенного 95%-ный доверительный интервал для рассматриваемой задачи в предположении, что $\pi = 0{,}4$, $n = 64$, $S_{\bar{s}} = \sqrt{\frac{0{,}4 \cdot 0{,}6}{64}} = 0{,}0612$, имеет следующие границы:

$$0{,}4 - 1{,}96 \cdot 0{,}0612 = 0{,}4 - 0{,}12 = 0{,}28;$$

$$0{,}4 + 0{,}12 = 0{,}52.$$

Иными словами, доля курящих студентов университета колеблется в пределах 28—52 %, что составляет от 2800 до 5200 человек.

Таким образом, для любой задачи с известной долей, полученной на основании выборки, всегда можно с определенным уровнем доверия найти процентные пределы колебания этой доли.

Например, из жителей населенного пункта, численность которого 10 000 человек, составлена выборка из 100 человек, 64 % которых одобрили некоторую федеральную программу. По этой доле можно оценить долю всех жителей, одобряющих эту программу. Она составит также 64 %. Однако колебания доли определяются из построения интервальной оценки. При этом в первую очередь определяют величину стандартного выборочного отклонения доли $S_{\bar{s}}$, а затем находят левую и правую границы интервала.

В данном случае $S_{\bar{s}} = \sqrt{\frac{0{,}64 \cdot 0{,}36}{100}} = 0{,}048$, на основании чего левая граница 95%-ного доверительного интервала равна $0{,}64 - 1{,}96 \cdot 0{,}048 = 0{,}64 - 0{,}094 = 0{,}546$, а правая граница — $0{,}64 + 0{,}094 = 0{,}734$. Это соответственно составляет 54,6 и 73,4 %.

Таким образом, от 54,6 до 73,4 % жителей поддерживают предложенную федеральную программу, т. е. от 5460 до 7340 человек.

Если полученный интервал кажется широким, его можно сузить, понизив уровень доверия, например, до 0,9. Если же при-

нять, что 95%-ный уровень доверия чрезвычайно важен, а интервал = const, необходимо уменьшить $S_{\bar{x}}$, для чего, в свою очередь, необходимо увеличить объем выборки, что следует из выражения $S_{\bar{x}} = \sqrt{\frac{P(1-P)}{n}}$.

Как было определено, точечная оценка разности средних арифметических двух выборок вычисляется как $\Delta_{\bar{x}_1-\bar{x}_2} = \bar{X}_1 - \bar{X}_2$, а интервальная оценка определяется по выражению (3.2) согласно формуле

$$(\Delta_{\bar{x}_1-\bar{x}_2} - Z \cdot \sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}}\ \Delta_{\bar{x}_1-\bar{x}_2} + Z \cdot \sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}}).$$

Аналогично точечная оценка разности долей двух генеральных совокупностей $\Delta_{p_1-p_2} = P_1 - P_2$, а интервальная оценка равна:

$$(\Delta_{p_1-p_2} - Z \cdot \sqrt{S^2_{p_1} + S^2_{p_2}},\ \Delta_{p_1-p_2} + Z \cdot \sqrt{S^2_{p_1} + S^2_{p_2}}).$$

Иными словами, средняя квадратичная ошибка разности долей определяется по выражению $\sqrt{S^2_{p_1} + S^2_{p_2}}$.

Например, в одном регионе страны доля P_1 жителей, одобряющих новый закон, равна 0,45 при средней квадратической ошибке $S_{\bar{x}} = 0{,}04$. В другом регионе доля P_2 жителей, поддерживающих закон, равна 0,55 при средней квадратической ошибке 0,03. Тогда при 95%-ном доверии пределы разности указанных долей определяются по выражению $(0{,}55 - 0{,}45) \pm 1{,}96\sqrt{0{,}04^2 + 0{,}03^2}$. Откуда получим 0,1 – 1,96 = 0,1 – 0,098 = 0,02, 0,1 + 0,098 = 0,198, что составляет 2,0 и 19,8 %.

3.2. Проверка гипотез относительно параметров

Под гипотезой (предположением) в статистике подразумевают всякое высказывание о значениях параметров генеральной совокупности или законе распределения случайной, которое проверяется на основании исследований выборки.

Например, если, опираясь на стандарты, предполагают, что средний срок службы телевизионных экранов μ составляет 2000 ч, затем берут случайную выборку, определяют статистику $\bar{X}$, сопоставляют ее с μ и устанавливают степень близости статистики $\bar{X}$ к параметру μ, тем самым проверяют гипотезу о правильности среднего срока службы телевизионных экранов.

В том случае, когда статистика $\bar{X}$ очень «близка» к параметру, считают предположение правильным и гипотезу принимают. Если же среднее значение выборки $\bar{X}$ значительно отличается от величины параметра μ, гипотеза отвергается и делается вывод о том, что принятое значение μ не соответствует действительности.

Таким образом, отличие проверки гипотез от оценки параметров генеральной совокупности состоит в том, что предположение (гипотеза) выдвигается до взятия выборки, а оценки находятся после взятия и исследования выборки.

В любом случае, принятия или непринятия гипотезы, на самом деле точное измерение значения параметра генеральной совокупности не производится. Речь идет об оценке величины разности между принятым значением параметра и полученным значением статистики. При этом истинность статистики не подвергается сомнению.

Идеальным доводом в пользу правильности гипотезы было бы обнаружение того факта, что разность между предполагаемым значением параметра генеральной совокупности и найденным значением статистики равна нулю. Поэтому проверяют разность, а гипотезу о проверке такой разности называют нулевой и обозначают H_0.

Таким образом, на самом деле проверяется предположение: разность между значением параметра и статистики равна нулю или нет? Противоположную гипотезу, разность не равна нулю или равна ему, называют альтернативной гипотезой и обозначают H_1.

В большинстве практических случаев нулевая гипотеза не подтверждается, т. е. разность между предполагаемым значением параметра генеральной совокупности и статистикой существует. Чтобы двигаться дальше, т. е. принять или отклонить гипотезу, необходимо установить, насколько значима эта разность. Если она значима, гипотеза отвергается, если нет — принимается.

Вопрос о значимости обычно ставится так: значимо ли отличие полученной величины от нуля?

Рассмотрим пример. Некоторое предприятие изготавливает подшипники, средний диаметр которых согласно стандарту $\mu = 1{,}461$ см. На основании выборки получен средний диаметр $\bar{X} = 1{,}435$ см. Таким образом, разность $\Delta_{\mu-\bar{X}} = 1{,}461 - 1{,}435 = 0{,}026$ см. Значимо ли отличие 0,026 от нуля?

Ответить на этот вопрос на основании только одной разности не представляется возможным. Поэтому для решения задачи используют понятие вероятности, рассуждая следующим образом.

Поскольку вероятность — это степень достоверности того, что наблюдаемое событие будет иметь желательный исход, а случайная величина примет ожидаемое значение, появление разности между статистикой и параметром генеральной совокупности явление нежелательное и вероятность этого появления должна быть очень мала. И наоборот, непоявление разности должно иметь высокую вероятность, свидетельствующую о том, что доверие к правильности принятого значения параметра генеральной совокупности велико.

Численно значения «малой» вероятности чаще всего принимают равными от 0,1 до 0,01. Эти значения называют соответственно 10%-ным и 1%-ным уровнями значимости. По существу они означают, что если разность между статистикой и принятым значением параметра генеральной совокупности столь велика, что вероятность случайного появления такой и большей разности равна 0,1 или 0,01 и меньше, т. е. разность считается значимой, и нулевая гипотеза отвергается при 0,1 или 0,01 уровнях значимости.

Чем больше разность, тем объективно она и более значима и, таким образом, степень непринятия нулевой гипотезы более высока. Например, если в предшествующим примере разность средних диаметров подшипников $\Delta_{\mu-\bar{X}} = 0{,}032$, то можно сказать, что такая разность скорее всего приведет к непринятию нулевой гипотезы H_0.

Теперь приведем графическую картину так называемых областей принятия и непринятия нулевой гипотезы H_0. Для этого рассмотрим пример проверки гипотезы относительно средней арифметической генеральной совокупности μ.

Как было показано ранее, для нормально распределенной случайной величины x вероятность ее попадания в интервалы $(\bar{X} - S, \bar{X} + S)$, $(\bar{X} - 1{,}64S, \bar{X} + 1{,}64S)$, $(\bar{X} - 1{,}96S, \bar{X} + 1{,}96S)$, $(\bar{X} - 2{,}58S, \bar{X} + 2{,}58S)$ равна 0,68, 0,9, 0,95 и 0,99, а вероятности непопадания в эти интервалы соответственно такие: 0,32, 0,1, 0,05, 0,01. Это означает, что, например, при вероятности $P = 0{,}95$, которой соответствует интервал $(\bar{X} - 1{,}96S, \bar{X} + 1{,}96S)$, для выборки объемом $n = 100$ девяносто пять значений случайной x попадут в интервал $(\bar{X} - 1{,}96S, \bar{X} + 1{,}96S)$, а пять ее значений окажутся за пределами интервала.

Когда речь шла об установлении доверительного интервала для средней арифметической генеральной совокупности μ, мы использовали более точное выражение $(\bar{X} \pm 1{,}96S_{\bar{x}})$, где $S_{\bar{x}}$ — стандартное выборочное отклонение, полученное на основании взятия множества выборок из генеральной совокупности. Поэтому, оценивая разность между средней выборки $\bar{X}$ и принятым значением генеральной совокупности μ, нужно пользоваться именно этим выражением. Тогда указанный интервал означает, что в 100 случаях средняя арифметическая будет попадать в этот интервал и только 5 из ее значений окажутся за его пределами. В свою очередь, это означает, что разность между статистикой и параметром μ настолько велика, что выходит за пределы интервала. Иными словами, она оказывается в зоне непринятия гипотезы.

На рис. 3.2 показаны границы интервалов для 5%-ного и 1%-ного уровней значимости относительно средней арифметической выборки.

Рис. 3.2. Границы интервалов уровней значимости:
а — 5%-ный уровень; *б* — 10%-ный уровень

На этом рисунке показаны также области принятия и непринятия гипотезы для 5%- и 1%-ного уровней значимости. Эти области установлены по определению: если разность между принятым параметром генеральной совокупности и статистикой настолько велика, что вероятность случайного появления такой и большей разности составляет менее 0,05 или 0,01, то эта разность значима и нулевая гипотеза отвергается.

Таким образом, нулевая гипотеза при 5%-ном уровне значимости отвергается всегда, если разность между параметром и статистикой столь велика, что такая и бóльшая разность будет наблюдаться в среднем в 5 из 100 случайных выборок. При 1%-ном уровне значимости нулевая гипотеза отвергается всегда, если разность столь велика, что она будет наблюдаться в одной из 100 случайных выборок. Отсюда следует, что непринятие нулевой гипотезы будет более частым при 5%-ном уровне значимости.

Объясним, как практически проверяется гипотеза относительно средней арифметической генеральной совокупности.

Прежде всего устанавливают уровень значимости, на основании чего определяют области принятия и непринятия гипотезы. Когда требуется весьма точная проверка средней арифметической генеральной совокупности μ, уровень значимости устанавливают высоким, т. е. 1%-ным. Когда же не требуется столь точная проверка, уровень значимости чаще всего принимают равным 0,05, т. е. 5%-ным.

Далее по выражению $\mu \pm 2{,}58S_{\bar{x}}$ или $\mu \pm 1{,}96S_{\bar{x}}$ находят критические пределы, отделяющие область принятия нулевой гипотезы от области ее непринятия. При этом значения $S_{\bar{x}}$ вычисляются по выражению $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}$ либо $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}}$, в зависимости от того, бесконечная или конечная генеральная совокупность и какой процент от N составляет объем выборки n.

По существу эти действия ничем не отличаются от определения границ доверительного интервала для средней арифметической генеральной совокупности. Разница состоит только в том, что интервал устанавливают не относительно точечной оценки средней арифметической генеральной совокупности, $\bar{X}$, а относительно предполагаемого значения средней генеральной совокупности μ.

При выборе интервала степень доверия, например, 0,95 указывает долю выборочных средних $\bar{X}$, равную 0,95 расположенных внутри интервала. При проверке гипотезы уровень значимости 0,05 определяет долю 0,05 выборочных средних $\bar{X}$, находящихся за пределами интервала.

После определения границ, разделяющих области принятия и непринятия нулевой гипотезы, выясняют, выходит или не выходит средняя выборки $\bar{X}$ за эти границы. Если выходит — гипотеза отвергается, если нет — принимается, т. е. значение μ считается правильным.

Рассмотрим пример. По утверждению руководства некоторой фирмы, средняя величина дебиторского счета составляет 5625 руб. Ревизор составляет случайную выборку из 50 счетов и устанавливает величину счета, равную 5250 руб. при стандартном отклонении 1050 руб.

Спрашивается, справедливо ли утверждение о том, что средняя величина дебиторского счета равна 5625 руб. при 5%- и 1%-ном уровнях значимости? Иными словами, будет принята или отклонена нулевая гипотеза при этих уровнях?

Для ответа на этот вопрос вначале определим границы, разделяющие области принятия и непринятия нулевой гипотезы. Поскольку границы находятся по выражению $\mu \mp 1{,}96 \cdot S_{\bar{x}}$, прежде всего вычислим значение $S_{\bar{x}}$. Так как количество элементов генеральной совокупности N не указано и $S = 1050$, $n = 50$, то $S_{\bar{x}} = \frac{S}{\sqrt{n-1}} = \frac{1050}{\sqrt{50-1}} = \frac{1050}{7} = 150$ руб. Откуда левая граница, вычисляемая по выражению $\mu - 1{,}96S_{\bar{x}}$, равна $5625 - 1{,}96 \cdot 150 = 5625 - 294 = 5331$ руб., а правая граница равна $5625 + 1{,}96 \times 150 = 5625 + 294 = 5919$ руб.

Таким образом, средняя арифметическая выборки $\bar{X} = 5250$ оказывается слева границы, равной 5361. Следовательно, нулевая гипотеза при 5%-ном уровне должна быть отвергнута и утверждение о средней величине дебиторского счета можно считать неправильным.

Для 1%-ного уровня значимости границы находятся по выражению $\mu \mp 2{,}58 \cdot S_{\bar{x}}$. Поэтому левая граница равна $5625 - 2{,}58 \times 150 = 5625 - 387 = 5238$ руб., а правая $5625 + 387 = 6012$ руб. В результате средняя арифметическая выборки $\bar{X} = 5250$ оказы-

вается внутри интервала (5238, 6012) и, следовательно, нулевая гипотеза при 1%-ном уровне значимости принимается. Это значит, что утверждение о средней величине дебиторского счета фирмы нужно считать правильным.

Различают двустороннюю и одностороннюю проверку нулевой гипотезы. Если нулевая гипотеза формулируется так, что предполагается проверка разности на ноль, такая проверка является двусторонней. Если же гипотеза предполагает проверку того, что статистика только больше или только меньше принятого параметра генеральной совокупности, такая проверка называется односторонней.

Например, если проверяется нулевая гипотеза относительно того, значимо или незначимо отклонение среднего диаметра подшипников $\bar{X}$ от стандартного μ, — это двусторонняя проверка. Когда же проверяется, значимо или незначимо превышение среднего диаметра подшипников над стандартным, мы имеем дело с односторонней проверкой.

Для двусторонней и односторонней проверок критические пределы, определяющие области принятия и непринятия гипотез, различны. Это показано на рис. 3.3 для 5%-ного уровня значимости.

Принятие или непринятие гипотезы всегда сопряжено с возможностью допущения ошибки. Различают ошибки первого и второго рода. Под ошибкой первого рода подразумевают ошибку, которую совершают, отвергая правильную гипотезу. Ошибка второго рода порождается принятием неверной гипотезы.

Фактически рассмотренные 5%-ные и 1%-ные уровни значимости можно толковать как значения вероятности 0,05 и 0,01 допустить ошибку первого рода. Чем ниже эти уровни, тем меньше эта вероятность, тем больше доверие к принятому значению параметра генеральной совокупности.

Таким образом, уровень значимости определяет степень риска (вероятность) допустить ошибку первого рода.

На практике вместо того, чтобы при проверке гипотез вычислять критические пределы, разделяющие области принятия и непринятия гипотез в натуральных величинах по формулам, например, $\mu \pm 1{,}96S_{\bar{x}}$, $\mu \pm 2{,}58S_{\bar{x}}$, используют критические пределы, выраженные в относительных величинах $Z = \dfrac{\mu - \bar{X}}{S_{\bar{x}}}$.

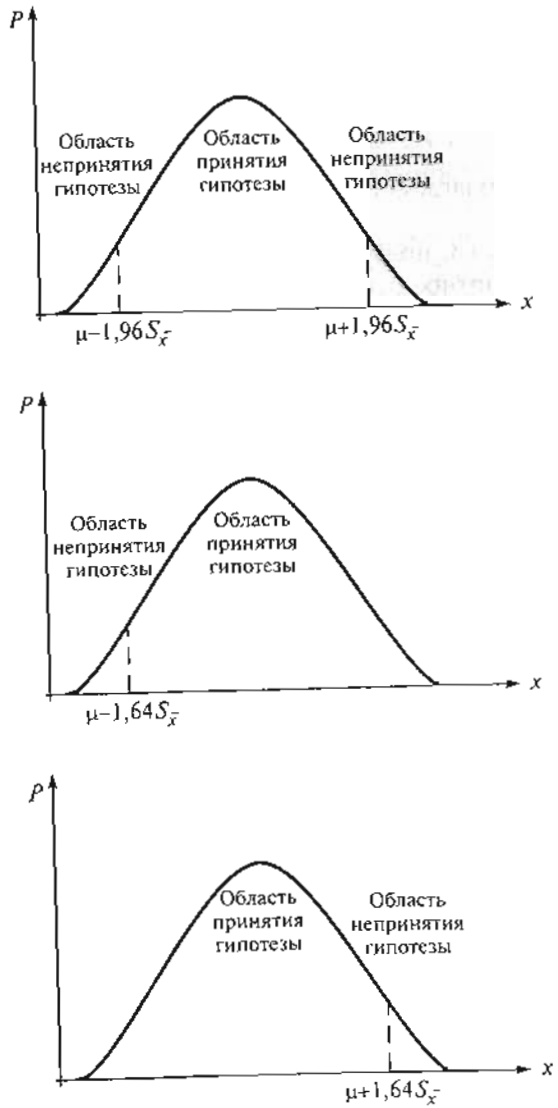

Рис. 3.3. Двусторонняя и односторонняя проверки гипотезы

Тогда для принятия или отклонения гипотезы действия сводятся к вычислению Z и сравнению его с 1,96 для 5%-ного либо с 2,58 для 1%-ного уровня значимости. Это даст возможность

проверять гипотезы относительно любых параметров, используя одну и ту же формулу

$$Z = \frac{P_r - S_i}{S_{\bar{S}}}, \qquad (3.4)$$

где P_r — предполагаемое значение параметра генеральной совокупности;

S_i – статистика, полученная на основании выборки;

$S_{\bar{S}}$ — стандартное отклонение статистики.

При этом для односторонней проверки гипотезы 5%-ного уровня значимости $Z = 1{,}64$ или $-1{,}64$, для 1%-ного $Z = 2{,}33$ либо $-2{,}33$.

Рассмотрим пример. Поставщик карманных фонарей утверждает, что средний срок службы батареи равен 1100 ч. Для выборки из 37 шт. фонарей средний срок службы оказался равным 1000 ч, а стандартное отклонение составило 240 ч.

Требуется проверить нулевую гипотезу 5%-ного уровня значимости относительно указанного срока службы фонарей — 1100 ч, т. е. установить, насколько достоверна эта информация.

Согласно формуле $Z = \frac{\mu - \bar{X}}{S_{\bar{x}}}$ предварительно по выражению $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}$ необходимо вычислить значение $S_{\bar{x}}$. Оно равно $\frac{240}{\sqrt{37-1}} = \frac{240}{6} = 40$. Тогда $Z = \frac{1100 - 1000}{40} = \frac{100}{40} = 2{,}5$. Так как $Z = 2{,}5 > 1{,}96 = Z_{кр}$, нулевая гипотеза отвергается, и, таким образом, срок службы фонарей, указанный поставщиком, сомнителен.

Обычно покупатель продукции, для которой указывается срок службы, большое внимание обращает на то, что указанный срок может оказаться меньше гарантированного.

Учитывая это, убедимся в том, что для данных предшествующей задачи срок службы фонарей при 5%-ном уровне значимости будет ниже 1100 ч.

В данном случае речь идет об односторонней проверке. Значение Z при 5%-ном уровне значимости для такой проверки равно $\pm 1{,}64$. Поскольку ранее вычисленное значение $Z = 2{,}5 > 1{,}64$,

то нулевая гипотеза отвергается. Поэтому, действительно, срок службы фонарей, рекламируемый продавцом, завышен.

На практике довольно часто приходится решать задачи сравнения средних генеральных совокупностей на основе средних арифметических двух выборок $\bar{X}$, $\bar{Y}$, взятых из одной генеральной совокупности или из разных совокупностей. Если средние значения $\bar{X}$, $\bar{Y}$ различны, возникает вопрос, случайное ли это различие, или оно закономерно. Если различие закономерно, то оно определяется влиянием определенного фактора.

Например, два обрабатывающих центра в определенном отрезке времени изготавливали втулки одинакового диаметра. Случайные выборки втулок показали, что их средние диаметры $\bar{X}$, $\bar{Y}$ различны. Необходимо установить, закономерно ли это различие, или оно случайно, т. е. определяется условиями проведения эксперимента?

Для этого требуется проверка нулевой гипотезы о равенстве средних $\bar{X} = \bar{Y}$ двух генеральных совокупностей. Так как общая формула вычисления критических пределов при проверке нулевой гипотезы имеет вид $Z = \frac{P_r - S_i}{S_{\bar{S}}}$ (выражение 3.4), для сравнения средних получим:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{\bar{X}-\bar{Y}}}, \qquad (3.5)$$

где μ_1, μ_2 — средние генеральной совокупности; $S_{\bar{X}-\bar{Y}}$ — стандартное выборочное отклонение разности, вычисляемое по формуле $S_{\bar{X}-\bar{Y}} = \sqrt{S_{\bar{X}}^2 + S_{\bar{Y}}^2}$.

Предположим, утверждают, что средняя зарплата преподавателя МГУ превышает среднюю зарплату преподавателя Курского государственного университета (КГУ) не больше чем на 15 000 руб. Проверим, правдиво ли это высказывание.

Для этого сформируем две случайные выборки: 54 преподавателя МГУ и 42 преподавателя КГУ. Оказалось, что среднее арифметическое выборок $\bar{X} = 51\,000$ руб. и $\bar{Y} = 35\,000$ руб., а стандартные отклонения $S_1 = 8100$ руб. и $S_2 = 7600$ руб.

При этих условиях и уровне значимости $a = 0{,}1$ проверим нулевую гипотезу, действительно ли $\mu_1 - \mu_2 \le 15\,000$, где $\mu_1 = 51\,000$, а $\mu_2 = 35\,000$ — зарплаты преподавателей МГУ и КГУ.

Сначала определим $S_{\bar{X}} = \frac{S_1}{\sqrt{n_1 - 1}}$, $S_{\bar{Y}} = \frac{S_2}{\sqrt{n_2 - 1}}$. В результате получим:

$$S_{\bar{X}} = \frac{8100}{\sqrt{54-1}} = \frac{8100}{\sqrt{53}} = 1112{,}6; \quad S_{\bar{Y}} = \frac{7600}{\sqrt{42-1}} = \frac{7600}{\sqrt{41}} = 1186{,}9.$$

Теперь вычислим $S_{\bar{X}-\bar{Y}} = \sqrt{(1112{,}6)^2 + (1186{,}9)^2} = 1626{,}8$.

Далее по формуле (3.5) найдем:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{\bar{X}-\bar{Y}}} = \frac{51\,000 - 35\,000 - 15\,000}{1626{,}8} = \frac{1000}{1626{,}8} = 0{,}615.$$

Критическое значение $Z_{кр}$ при односторонней проверке и $a = 0{,}1$ равно 1,28. Так как $Z = 0{,}615 < 1{,}28 = Z_{кр}$, оно попадает в область принятия нулевой гипотезы. Поэтому, действительно, с вероятностью 0,9 разность зарплат преподавателей не превышает 15 000 руб.

В том случае, когда выборки берутся из одной и той же генеральной совокупности, $\mu_1 = \mu_2$, формула для вычисления Z упрощается. Она будет иметь такой вид:

$$Z = \frac{(\bar{X} - \bar{Y})}{S_{\bar{X}-\bar{Y}}}. \qquad (3.6)$$

Рассмотрим пример. Некоторый станок изготовляет подшипники. Их средний диаметр $\bar{X}$, найденный по выборке $n = 65$, равен 0,609 см, а стандартное отклонение $S_1 = 0{,}05$ см. Выборка объема $n = 65$, взятая из подшипников, изготовленных на этом же станке в следующую смену, показала, что средний их диаметр $\bar{X} = 0{,}635$ см при $S_2 = 0{,}101$ см.

Необходимо установить, случайно ли это различие или закономерно.

Гипотезу о равенстве средних диаметров подшипников для двух выборок, взятых из одной и той же генеральной совокупности, проверим при 5%-ном уровне значимости. Для этого определим критические пределы, используя формулу (3.4), предварительно вычислив стандартное выборочное отклонение для разности двух средних $S_{\bar{X}-\bar{Y}}$.

Получим:

$$S_{\bar{X}} = \frac{S_1}{\sqrt{n_1 - 1}} = \frac{0{,}05}{\sqrt{65-1}} = \frac{0{,}05}{8} = 0{.}006;$$

$$S_{\bar{Y}} = \frac{S_2}{\sqrt{n_2 - 1}} = \frac{0{,}101}{\sqrt{65-1}} = \frac{0{,}101}{8} = 0{,}013;$$

$$S_{\bar{X}-\bar{Y}} = \sqrt{S_{\bar{X}}^2 + S_{\bar{Y}}^2} =$$

$$= \sqrt{(0{,}006)^2 + (0{,}013)^2} = \sqrt{0{,}000036 + 0{,}000169} = \sqrt{0{,}000205} = 0{,}0143;$$

$$Z = \frac{(\bar{X} - \bar{Y})}{S_{\bar{X}-\bar{Y}}} = \frac{0{,}609 - 0{,}635}{0{,}0143} = \frac{-0{,}026}{0{,}0143} = -1{,}8.$$

Так как при 5%-ном уровне значимости критические пределы $Z = \pm 1{,}96$, то $-1{,}8 > -1{,}96$, гипотеза принимается, т. е. разность средних диаметров не закономерна, а порождается случайными обстоятельствами проведения эксперимента.

Проверка гипотез может быть осуществлена и для тех случаев, когда рассматриваемые случайные величины дискретны.

Рассмотрим такой пример. Монета подбрасывается пять раз, и все пять раз выпадает орел. Можно ли при 5%-ном уровне значимости утверждать, что монета несимметрична? Иными словами, необходимо установить, значима или незначима выдвинутая гипотеза при уровне 0,05.

Чтобы решить эту задачу, необходимо построить закон распределения вероятности выпадения орла. В свою очередь, для этого необходимо найти коэффициенты разложения бинома Ньютона $(a + b)^5$, для чего можно воспользоваться формулой (2.8), которая для данного случая будет иметь такой вид:

$$P(x) = C_5^x P^x \cdot (1 - P)^{5-x}, \quad x = 1, 2, 3, 4, 5,$$

где x — число выпадений орла;

P — вероятность его выпадений.

Выполнив необходимые вычисления в предположении, что $P = \frac{1}{2}$, получим следующее распределение вероятностей (рис. 3.4).

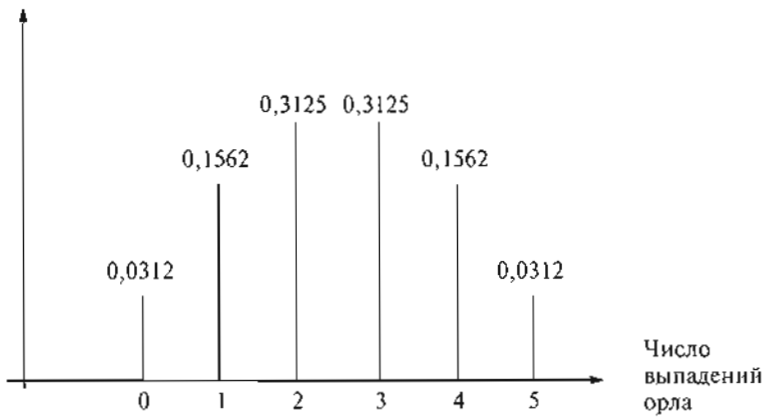

Рис. 3.4. Распределение вероятности выпадения орла

На основании этого распределения можно сказать, что гипотеза не значима, так как 0,0312 + 0,0312 = 0,0624 > 0,05.

В более сложных случаях проверки гипотез для дискретных случайных величин вместо биномиального используют нормальное распределение вероятностей. Как и для интервальных оценок, замена дискретного распределения на нормальное осуществима в случае, если объем выборки $n \geq 50$ и $n \cdot a \geq 5$, $n \cdot b \geq 5$.

Рассмотрим пример. На основании случайного опроса было установлено, что 4 студента из 10 некоторого университета постоянно курят. Таким образом, доля выборки составила $P = 0,4$. В результате этого был сделан вывод о том, что эта же доля курящих свойственна всем студентам, составляющим генеральную совокупность, т. е. $\pi = 0,4$.

Средняя квадратическая ошибка доли генеральной совокупности получена на основании формулы $S_{\overline{s}} = \sqrt{\frac{P(1-P)}{n}}$ и равна $\sqrt{\frac{0,4 \cdot 0,6}{64}} = 0,0612$.

Некоторые статистики полученный результат подвергли сомнению и провели свой эксперимент, взяв случайную выборку из 64 студентов. Оказалось, что 40 студентов из 64 курят. Таким образом, доля P составила 40/64 = 0,625.

Требуется проверить гипотезу, действительно ли число курящих студентов университета больше 40 %.

В данном случае речь идет об односторонней проверке, поэтому $Z_{кр} = 1,64$. На основании формулы $Z = \frac{(P - \pi)}{S_{\overline{s}}}$ получаем

$$Z = \frac{0,625 - 0,4}{0,0612} = 3,68 > 1,64.$$

Таким образом, выдвинутая гипотеза не принимается.

Весьма часто, особенно в различных программных средствах, предназначенных для решения задач статистики, используется понятие *наблюдаемого уровня значимости*. Под этим уровнем понимают наименьший уровень значимости, при котором отклоняется нулевая гипотеза при том условии, что она истинна.

Например, если для $a = 0,05$, $Z_{кр} = 1,96$, а полученное значение $Z = 1,7$ — гипотеза принимается. Когда же, например $Z = 2,05$, при этом уровне значимости она отклоняется.

Для первого случая наблюдаемый уровень значимости — это вероятность того, что $Z > 1,7$, т. е. $P(Z > +1,7) = 1 - P(Z \leq 1,7)$. Для второго случая $P(Z > +2,05) = 1 - P(Z \leq +2,05)$.

Так как согласно Z-распределению $1 - P(Z \leq 1,7) = 0,0446$ и это значение меньше $a = 0,05$, то? учитывая, что для этого случая гипотеза принимается, делаем следующий вывод.

Если наблюдаемый уровень значимости больше принятого уровня a, гипотеза принимается. Как только он станет равным или меньше его, гипотеза отклоняется. При этом следует помнить, для двусторонней проверки гипотезы вероятность $P(Z >$ число$)$ должна быть удвоена.

Проверка гипотез может быть осуществлена и для разности долей. В этом случае, когда $n \geq 50$, для вычисления Z используется выражение

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{S_{p_1 - p_2}},$$

где π_1, π_2 — доли генеральной совокупностей;

P_1, P_2 — доли выборок.

Когда выборки берутся из одной и той же генеральной совокупности, тогда $\pi_1 = \pi_2$ и Z вычисляется по более простой формуле $Z = \frac{P_1 - P_2}{S_{p_1 - p_2}}$.

В свою очередь, согласно изложенному при рассмотрении точечных оценок долей S и их разностей

$$S_{p_1 - p_2} = \sqrt{S_{p1}^2 + S_{p2}^2}; \quad S_{p_1} = \sqrt{\frac{P_1(1 - P_1)}{n_1}}; \quad S_{p2} = \sqrt{\frac{P_2(1 - P_2)}{n_2}}.$$

Когда же n мало, т. е. меньше 30, следует использовать биномиальное распределение, что в общем усложняет решение задачи проверки гипотез.

Рассмотрим пример. При изучении мнения покупателей относительно внешнего вида шубы всех покупателей, проявивших интерес к указанному изделию, разбили на группы согласно их семейному доходу. Затем из каждой группы взяли случайные выборки $n_1 = 64$, $n_2 = 64$ и задали такой вопрос: «Положительно ли вы оцениваете изделие?» Оказалось, что 45 % первой выборки и 55 % второй выборки ответили «да».

Требуется проверить нулевую гипотезу 5%-ного уровня значимости относительного того, что различие в мнении покупателей не случайно.

Вначале вычислим $S_{p_1} = \sqrt{\frac{0{,}45(1 - 0{,}45)}{64}} = 0{,}062;$

$$S_{p_2} = \sqrt{\frac{0{,}55(1 - 0{,}55)}{64}} = 0{,}062.$$

Далее получим $S_{p_1 - p_2} = \sqrt{0{,}062^2 + 0{,}062^2} = 0{,}088.$

Таким образом, $Z = \frac{0{,}45 - 0{,}55}{0{,}088} = -1{,}136.$

Так как $Z = -1{,}136 > -1{,}96 = Z_{кр}$, гипотеза принимается, т. е. мнение покупателей относительно заданного вопроса подтверждается.

3.3. Проверка гипотез по критерию хи-квадрат

Нами были рассмотрены правила проверки гипотез относительно параметров генеральной совокупности. Теперь расскажем о том, как проверять гипотезы относительно законов распределения частот случайных величин. Эта задача практически становится чрезвычайно важной в тех случаях, когда речь идет о замене того или иного закона нормальным распределением, поскольку большинство практических приемов аналитической статистики опирается на этот закон.

Распределение ХИ-квадрат (χ^2) Пирсона представляет собой распределение вероятностей квадратов k независимых случайных величин, каждая из которых распределена по нормальному закону с нулевым средним арифметическим и дисперсией, равной единице. Геометрически распределение представляет собой семейство кривых для $k = 1, 2, 3, \ldots$. Каждая кривая асимметрична, с положительной асимметрией, и, только начиная с $k = 10$, соответствующая кривая отражает закон распределения, близкий к нормальному.

По своей природе распределение χ^2 является распределением дискретных величин.

То обстоятельство, что, с одной стороны, распределение Пирсона является суммой квадратов случайных, а с другой — при помощи суммы квадратов измеряются отклонения, ошибки и, в частности, дисперсия, дает основание использовать это распределение для проверки отклонения эмпирических и теоретических частот и, таким образом, сопоставлять законы распределения.

В вычислительном отношении проверка по критерию χ^2 позволяет установить значимость разностей между наблюдаемыми и прогнозируемыми частотами.

В практических условиях возникает два случая: 1) известно некоторое распределение частот и требуется эмпирически с определенной вероятностью подтвердить или отвергнуть закон распределения; 2) требуется подтвердить или отвергнуть опять-таки с наперед заданной вероятностью, что полученное в результате эксперимента или наблюдения распределение отвечает известному закону.

Обе эти задачи носят название проверки на согласие или адекватность распределений.

Рассмотрим примеры. Новый фильм, который скоро должен появиться на экранах, оценивается специалистами по пяти категориям: 1 — прекрасный; 2 — между прекрасным и посредственным; 3 — посредственный; 4 — ниже посредственного; 5 — плохой. Соответствующие показатели относительных частот, полученных на основании опроса 100 специалистов, приведены в табл. 3.1 (колонка 2).

Таблица 3.1. Оценки качества фильма

Категории	Процент	Кол-во опрошенных зрителей, f_{in}	Объем выборки	Прогнозируемые частоты, f_{ip}	$f_{in} - f_{ip}$	$(f_{in} - f_{ip})^2$	$\frac{(f_{in} - f_{ip})^2}{f_p}$
1	40	145	400	0,4 · 400 = 160	–15	225	1,41
2	30	128	400	0,3 · 400 = 120	8	64	0,53
3	20	73	400	0,2 · 400 = 80	–7	49	0,61
4	5	32	400	0,05 · 400 = 20	12	144	7,20
5	5	22	400	0,05 · 400 = 20	2	4	0,20

После премьеры фильма был произведен опрос 400 зрителей, просмотревших фильм. Их количества по категориям приведены в колонке 3. По существу они представляют собой частоты, полученные на основании выборки $n = 400$, т. е. наблюдаемые частоты.

Для того чтобы получить прогнозируемые частоты, приведенные к объему выборки, необходимо прогнозируемую частоту каждой категории, т. е. элементы колонки 2, умножить на $n = 400$ и разделить на 100. Полученные данные представлены в колонке 5.

Теперь можно перейти к вычислению значения χ^2. Для этого применяется следующая формула:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{in} - f_{ip})^2}{f_{ip}}, \qquad (3.5)$$

где k — число групп частот;

f_{in}, f_{ip}, $i = 1, 2, 3, \ldots$ — наблюдаемые и прогнозируемые частоты.

Дальнейший расчет χ^2 представлен колонками 6—8. Полученное значение $\chi^2 = 9,95$ — сумма чисел последней колонки.

Чтобы использовать это значение для оценки законов совпадения распределения частот, необходимо сформулировать нулевую и альтернативную гипотезы. Полученные в результате опроса кинозрителей частоты при заданном уровне значимости соответствуют прогнозируемым частотам или указанные частоты

далеки от наблюдаемых. Кроме того, необходимо вычислить число степеней свободы $d = k - 1$ для определения соответствующей кривой распределения χ^2.

В нашем случае k равно числу групп частот. Следовательно, $d = 5 - 1 = 4$. После этого для заданного уровня значимости, например $\alpha = 0,05$, т. е. значения вероятности, при которой нулевая гипотеза отвергается, необходимо по специальной таблице либо при помощи специальной программы найти критическое значение $\chi^2_{кр}$, соответствующее этой вероятности. В данном случае для $d = 4$ и $\alpha = 0,05$ оно равно 7,779. Если найденное значение $\chi^2 \geq \chi^2_{кр}$, нулевая гипотеза отклоняется, так как вероятность для $\chi^2_{кр} \leq 0,05$. В противном случае гипотеза принимается.

Ввиду того что $\chi^2 = 9,95 > 7,779 = \chi^2_{кр}$, нулевая гипотеза отклоняется, мы делаем вывод о том, что при уровне значимости 0,05 фактическое распределение частот оценки фильма отличается от прогнозируемого специалистами. Данная ситуация изображена на рис. 3.5.

Рис. 3.5. Проверка гипотезы по критерию χ^2

Таким образом, нулевая гипотеза всегда будет отвергнута, если $\chi^2 \geq \chi^2_{кр}$ для любого уровня значимости. Определение $\chi^2_{кр}$ по таблице в настоящее время практически не производится. Для этого в пакете **Анализ данных** предусмотрена специальная программа, о которой мы расскажем в следующем параграфе.

Предположим, 20 раз подбрасываются две монеты и в каждом подбрасывании фиксируется число выпадений орла. Полученные данные следующие: невыпадение орла — 4 случая, одно

выпадение — 8 случаев, два выпадения -- 8 случаев. Известно, что такой опыт отражает биномиальное распределение вероятностей числа выпадений орла.

По полученным частотам при заданном уровне значимости требуется подтвердить или отвергнуть гипотезу о том, что они отражают указанное теоретическое распределение.

Для решения этой задачи на основании разложения бинома Ньютона $(a + b)^2$ при $a = 1/2$, $b = 1/2$ необходимо определить значения коэффициентов разложения — вероятности успехов. Затем по этим вероятностям найти теоретические частоты при 20 подбрасываниях монет. После этого вычислить значение χ^2.

Разложение $(a + b)^2 = a^2 + 2ab + b^2 = 1/4 + 1/2 + 1/4$.

Теоретические частоты равны $1/4 \cdot 20 = 5$, $1/2 \cdot 20 = 10$, $1/4 \cdot 20 = 5$.

Используя формулу (3.5), получим:

$$\chi^2 = \frac{(4-5)^2}{5} + \frac{(8-10)^2}{10} + \frac{(8-5)^2}{5} = \frac{(-1)^2}{5} + \frac{(-2)^2}{10} + \frac{3^2}{5} = 2{,}4.$$

Для определения χ^2 необходимо указать число степеней свободы, $d = k - 1$. В данном случае оно равно $3 - 1 = 2$. Критическое значение $\chi^2_{кр}$ для $d = 2$ и уровня значимости 0,05 равно 5,99. Поскольку $\chi^2 = 2{,}4 < 5{,}99 = \chi^2_{кр}$, нулевая гипотеза незначима и она не отвергается.

Таким образом, с вероятностью 0,95 полученное эмпирическое распределение соответствует биномиальному закону распределения.

Аналогичный подход может быть использован и при проверке согласия эмпирических частот с нормальным законом распределения. В этом случае теоретически частоты определяются на основании значений плотности вероятности теоретического распределения по формуле (2.1) или при помощи функции **Нормрасп**(x, $\bar{X}$, σ, 0) из пакета **Анализ данных**.

Рассмотрим пример, взятый из [5].

Предположим, на основании измерения роста 500 студентов некоторого университета и группировки данных получены результаты, представленные в колонках 2—4 табл. 3.2: рост студентов по группам, наблюдаемая частота в группе f_{in}, метка группы X_s.

Требуется подтвердить или опровергнуть гипотезу о том, что при уровне значимости $\alpha = 0{,}05$ эмпирическое распределение частот соответствует нормальному.

Таблица 3.2. **Проверка гипотезы о нормальном распределении частот**

№	Рост студентов по группам	f_{in}	X_s	$(X_s - \bar{X})^2$	$f(X_s, \bar{X}, \sigma)$	f_{ip}	$\frac{(f_{in} - f_{ip})^2}{f_{ip}}$
1	162—166	5	163	211,06	0,00323	6,47	0,334
2	166—170	33	168	110,84	0,01371	27,42	1,136
3	170—174	70	172	42,61	0,03665	73,30	0,148
4	174—178	132	176	6,39	0,06177	123,54	0,579
5	178—182	119	180	2,17	0,06565	131,29	1,150
6	182—186	87	184	29,94	0,04339	87,98	0,011
7	186—190	42	188	89,72	0,01859	37,18	0,624
8	190—194	12	192	181,49	0,00495	9,91	0,441

Для того чтобы решить поставленную задачу, необходимо применить формулу (3.5) и определить значение χ^2. В свою очередь, для этого требуется по каждой группе найти прогнозируемую частоту f_{ip}.

Такие частоты вычисляются на основании умножения значений плотности вероятности нормального распределения $f(X_s, \bar{X}, \sigma)$, соответствующих X_s, на длину интервала, равную 4, и объем выборки $n = 500$.

Таким образом, процедура сводится к вычислению значений плотности вероятности $f(X_s, \bar{X}, \sigma)$. Для этого используется формула

$$f(X_s, \bar{X}, \sigma) = \frac{e^{-\frac{(X_s - \bar{X})^2}{2\sigma^2}}}{\sigma \cdot \sqrt{2\pi}}$$

— измененное выражение (2.11), в котором вместо варианты x используется метка группы X_s, а вместо средней арифметической μ — генеральной совокупности — средняя арифметическая выборки $\bar{X}$.

В свою очередь, для вычисления значений $\bar{X}$ и σ применяются выражения (1.2), (1.11), в которых вместо x_{ls}, f_l используются соответственно X_s, f_{in}. В результате на основании (1.2), (1.11) получаем $\bar{X} = 178{,}528$, $\sigma = 5{,}98$.

Используя эти значения, последовательно заполняем колонки 5 и 6 таблицы. Затем по выражению $f(X_s, \bar{X}, \sigma) \cdot 4 \cdot 500$ находим значения прогнозируемой частоты f_{ip} и далее — отношения $\frac{(f_{in} - f_{ip})^2}{f_{ip}}$, представленные колонкой 8. Суммируя значения колонки 8, окончательно получаем $\chi^2 = 4{,}423$.

При числе степеней свободы $d = k - 1 = 8 - 1 = 7$ и уровне значимости $\alpha = 0{,}05$ критическое значение $\chi^2_{кр} = 14{,}07$. Так как $\chi^2 = 4{,}33 < 14{,}07 = \chi^2_{кр}$, гипотеза принимается, т. е. различие между эмпирическим и теоретическим законами распределения частот при $\alpha = 0{,}05$ — незначимо.

Следовательно, рост студентов подчиняется нормальному закону распределения.

В том случае, когда для вычисления плотности вероятности использовалась бы функция **Нормрасп**, необходимо было бы восемь раз обратиться к этой функции. Тем не менее этот расчет, безусловно, оказался бы проще.

Заканчивая этот параграф, необходимо отметить, что применение распределения χ^2 для выявления согласия между эмпирическими и прогнозируемыми распределениями частот имеет некоторые ограничения. Они начинают играть существенную роль, когда эмпирические частоты групп данных очень низкие, а именно меньше 5. В этом случае сказывается дискретность χ^2-распределения, что вносит ошибки при вычислении конкретного его значения. Поэтому на практике группы с низкой частотой рекомендуется объединять в одну группу.

Например, в табл. 3.3 приведены количества ошибок и их частоты на 100 писем, которые печатает секретарь одной из фирм.

Таблица 3.3. Частоты ошибок на 100 писем

Число ошибок	0	1	2	3	4	5
Ожидаемая частота	65	15	8	5	4	3

В связи с тем что частоты четырех и пяти ошибок меньше 5, группы данных 4 и 5 следует объединить в одну группу 4 с соответствующей частотой, большей 5.

Теперь более подробно объясним, что в статистике понимают под числом степеней свободы. По определению — это число элементов, которые могут свободно изменяться.

Например, пусть x_1 и x_2 — значения наблюдений со средним $\bar{X} = \frac{(x_1 - x_2)}{2} = 7$. Возможные значения x_1 и x_2 при этом средним — все те числа, которые в сумме дают $x_1 + x_2 = 14$. Если x_1 положить равным 11, то, исходя из этой суммы, x_2 может быть равным только 3, т. е. в этом случае мы можем изменять только одну переменную и, следовательно, число степеней свободы равно 1.

Таким образом, если выборка состоит из двух элементов со значениями x_1, x_2 и статистика для этой выборки вычислена как $\bar{X} = \frac{x_1 + x_2}{2} = 7$, число элементов, которые могут изменяться, равно 1. Если же выборка состоит из n элементов и значения средней вычислено для определения интервальной оценки и проверки гипотезы, число степеней свободы $d = n - 1$. В тех случаях, когда осуществляется проверка по критерию χ^2, число степеней свободы $d = r - 1$, где r — число групп частот.

3.4. Интервальные оценки и проверка гипотез для малых выборок

Рассмотренные методы вычисления интервальных оценок и проверки гипотез опирались на результаты проявления центральной предельной теоремы:

1) выборочные средние $\bar{X}$ для выборок объемом $n \geq 30$ даже в том случае, когда исходная совокупность имеет любой закон распределения, распределяются по нормальному закону;

2) они имеют тенденцию концентрироваться вокруг среднего μ генеральной совокупности;

3) стандартные ошибки средних $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}$ уменьшаются при увеличении выборки.

Эти особенности давали полное право при проведении расчетов использовать нормальный закон распределения, и в частности Z-распределение, где Z определялось по формуле $Z = \frac{\overline{X} - \mu}{\sigma_{\overline{x}}}$ либо $Z = \frac{\overline{X} - \mu}{S_{\overline{x}}}$.

Однако в тех случаях, когда $n < 30$, т. е. при малых выборках, которые преобладают при статистических исследованиях в промышленности, сельском хозяйстве и других областях, влияние центральной предельной теоремы с уменьшением n ослабевает. Вследствие этого нормальное распределение для вычисления интервальных оценок и проверки гипотез можно использовать лишь при определенных допущениях.

Когда выборка одновременно и генеральная совокупность, для которой известно стандартное отклонение σ, можно использовать Z-распределение, определяемое по формуле $Z = \frac{\overline{X} - \mu}{\sigma_{\overline{x}}}$. В этом случае оно сохраняет нормальный закон распределения, так как $\sigma = \text{const}$ и $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \text{const}$ при постоянном n, равном количеству элементов генеральной совокупности.

Таким образом, в этом случае для вычисления интервальных оценок и проверки гипотез можно использовать приведенные выше формулы (3.1)—(3.5).

Когда же выборка имеет малый объем, т. е. $n < 30$, и она не равна генеральной совокупности, а σ неизвестно, Z-распределение, определяемое по выражению $Z = \frac{\mu - \overline{X}}{S_{\overline{x}}}$, использовать нельзя. Объясняется это тем, что $S_{\overline{x}}$ вычисляют по формуле $S_{\overline{x}} = \frac{S}{\sqrt{n-1}}$, которая не обеспечивает $S_{\overline{x}} = \text{const}$ из-за разных значений S стандартных отклонений выборки.

Влияние величины выборки на характер распределения их средних при объемах выборки $n < 30$ было изучено У. Госсетом, публиковавшим результаты своих исследований под псевдонимом Стьюдент. Он установил, что для каждого n распределение носит симметричный характер, кривая распределения имеет колоколообразный вид, свойственный нормальному распределению, однако она имеет эксцесс, меньший нормального, и он уменьшается при уменьшении n. Когда же объем выборки n приближается к 30 и становится больше его, распределение средних приближается к нормальному.

На рис. 3.6 показаны кривые нормального распределения и распределения Стьюдента, семейство которых, определяемое различными значениями n, получило название t-распределения.

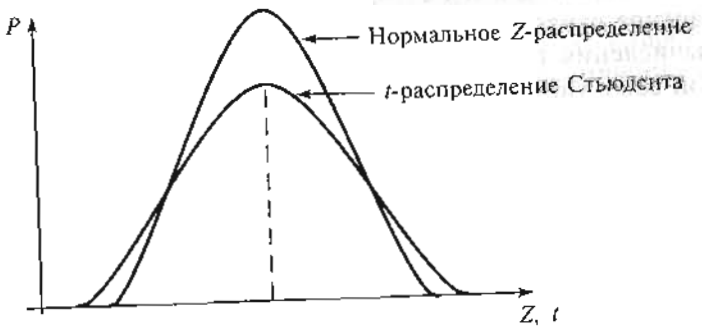

Рис. 3.6. Нормальное распределение Z и t-распределение

Распределение Стьюдента для конкретного n представляет собой распределение непрерывной случайной величины. По существу — это семейство распределений, представленное кривыми, зависящими от степеней свободы $k = n - 1$. Для нормированного t-распределения, вычисляемого по формуле $t = \frac{\mu - \overline{X}}{S_{\overline{x}}}$, и каждого k составлена таблица площадей (вероятностей), по которой всегда для любых t и k можно определить вероятность $P(x < t)$. Эти же возможности заложены и в пакете **Анализ данных** Excel Microsoft Office.

Сопоставляя кривые нормального Z-распределения и t-распределения Стьюдента, можно видеть, что t-распределение на «хвостах» имеет бóльшие площади, чем нормальное распределение. Объясняется это тем, что площади, охватываемые обеими кривыми, одинаковы и равны 1, а эксцесс t-распределения меньше, чем для нормального распределения вероятностей.

Поэтому, если интервальные оценки вычислять на основании t-распределения, интервал при данном уровне доверия, на-

пример 0,95, будет шире, чем при использовании нормального Z-распределения. При t-распределении необходимо сильнее отклоняться от средней μ, чтобы охватить ту же долю площади под кривой, которая соответствует нормальному Z-распределению.

По этой же причине при использовании t-распределения для проверки гипотезы критическое значение $t_{кр}$ при данном уровне значимости будет дальше от центра распределения μ, чем соответствующее значение $Z_{кр}$.

Таким образом, в тех случаях, когда объем выборки $n \geq 30$ и стандартное отклонение генеральной совокупности σ известно, для вычисления пределов интервальных оценок средней генеральной совокупности μ следует использовать выражение

$$\bar{X} \pm Z\sigma_{\bar{x}}, \tag{3.6}$$

определив предварительно $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Для оценки суммы всех элементов генеральной совокупности, как и раньше, можно использовать формулу

$$N(\bar{X} + Z\sigma_{\bar{x}}). \tag{3.7}$$

Для оценки разности двух средних $\bar{X}_1$, $\bar{X}_2$ — формулу

$$(\bar{X}_1 - \bar{X}_2) \pm Z\,\sigma_{\bar{x}_1 - \bar{x}_2}, \tag{3.8}$$

где $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}$.

Когда же $n \leq 30$ и стандартная ошибка $S_{\bar{X}}$ вычисляется на основании стандартных отклонений выборки $S_{\bar{X}} \frac{S}{\sqrt{n-1}}$, для вычисления пределов интервальных оценок средней генеральной совокупности необходимо использовать t-распределение Стьюдента. Поэтому формулы (3.6)—(3.8) будут иметь такой вид:

$$\bar{X} \pm t \cdot S_{\bar{x}}; \tag{3.9}$$

$$N(\bar{X} + tS_{\bar{x}}); \tag{3.10}$$

$$(\bar{X}_1 - \bar{X}_2) \pm t\,S_{\bar{x}_1 - \bar{x}_2}. \tag{3.11}$$

Причем при определении t в выражении (3.11) следует учитывать, что число степеней свободы k равно сумме чисел каждой выборки минус 2, т. е. $k = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$.

Рассмотрим примеры. Некоторое предприятие производит батареи для карманных фонарей. Установлено, что для выборки из 10 батарей ($n = 10$) средний срок службы $\bar{X}$ равен 10 ч, а стандартное отклонение генеральной совокупности σ — 3 ч.

Тогда границы интервалов при определенном уровне доверия следует вычислять по формуле (3.6).

Предположим, что для той же выборки из 10 батарей средний срок службы $\bar{X}$ равен 10 ч, а стандартное отклонение выборки S — 3 ч.

Тогда границы интервалов при определенном уровне доверия необходимо вычислять по выражению (3.9). При этом $S_{\bar{X}} = \frac{S}{\sqrt{n-1}} = \frac{3}{\sqrt{10-1}} = 1$, а число степеней свободы $k = n - 1 = 10 - 1 = 9$. По уровню доверия 0,95 и $k = 9$ находим t и вычисляем границы интервалов для среднего μ.

Если для этого случая необходимо оценить сумму элементов генеральной совокупности, например для $N = 1000$ шт. батарей, то следует использовать выражение (3.10). Учитывая, что границы интервала уже найдены, их просто необходимо умножить на 1000. Тем самым будут получены пределы суммарного срока службы 1000 батарей.

Предположим, необходимо оценить различие в качестве батарей, изготавливаемых двумя предприятиями. Для выборки из 17 батарей первого предприятия средний срок службы равен 22 ч при стандартном отклонении $S_1 = 6$ ч. Для выборки из 10 батарей второго предприятия средний срок службы равен 18 ч при стандартном отклонении $S_2 = 3$ ч. Тогда для получения оценки разности необходимо использовать выражение (3.11).

Получим $S_{\bar{x}_1} = \frac{S_1}{\sqrt{17-1}} = \frac{6}{4} = 1{,}5$, $S_{\bar{x}_2} = \frac{S_2}{\sqrt{10-1}} = \frac{3}{3} = 1$, так как $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}}$ то $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{1{,}5^2 + 1^2} = \sqrt{3{,}25} = 1{,}803$.

Учитывая, что $k = n_1 - 1 + n_2 - 1 = 17 - 1 + 10 - 1 = 25$, для уровня доверия 0,95 и $k = 25$ найдем t и определим разность по формуле (3.11).

В компьютерных программах проверки гипотез о разности средних вместо выражения для вычисления ошибки разности $S_{\bar{x}_1-\bar{x}_2} = \sqrt{S_{\bar{x}_1}^2 + S_{\bar{x}_2}^2}$ используются более точные формулы. Причем они различны для случаев равных и различных стандартных отклонений, т. е. различаются для $S_1 = S_2$ и $S_1 \neq S_2$.

Для случая $S_1 = S_2$ применяется выражение

$$S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Для случая $S_1 \neq S_2$ используется выражение $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$, а число степеней свободы определяется как $d = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{k}$, где

$$k = \frac{\left(\frac{S_1}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2}{n_2}\right)^2}{n_2-1}.$$

При проверке гипотез для выборок объемом $n \leq 30$, так же как и при вычислении интервальных оценок, используются и Z-распределение и t-распределение. Когда известны предполагаемое значение μ генеральной совокупности и σ, используется Z-распределение, по которому вычисляется критическое значение $Z_{кр} = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$. Когда же вместо σ мы располагаем стандартным отклонением выборки S, по формуле $t = \frac{\bar{X}-\mu}{S_{\bar{X}}}$ вычисляется значение t, которое далее сравнивается с $t_{кр}$, полученным по заданному количеству степеней свободы и уровню значимости α.

Предположим, согласно стандарту средний срок службы батареи равен 22 ч, а стандартное отклонение σ всех выпускаемых в данный момент батарей равно 2 ч. Для выборки 90 батарей средний срок службы оказался равным 20 ч.

В этом случае для проверки нулевой гипотезы об истинности μ следует использовать Z-распределение и сравнивать полученное Z с $Z_{кр}$, так как известно σ. Если же σ неизвестно, а получено стандартное отклонение выборки S, равное, например, 3 ч, и $\bar{X}$, равное 20 ч, необходимо использовать t-распределение, для чего вычислять t и сравнивать его с $t_{кр}$.

Ввиду того что объем выборки $n = 10$, $S_{\bar{X}} = \frac{S}{\sqrt{n-1}} = \frac{3}{\sqrt{9}} = 1$.

Получаем $t = \frac{\bar{X}-\mu}{S_{\bar{X}}} = \frac{20-22}{1} = -2$.

Критические значения $t_{кр}$ для уровня значимости 0,05 равны −2,262, +2,262. В связи с тем что t попадает в этот интервал, т. е. лежит в области принятия гипотезы, разность $\bar{X} - \mu$ незначима и можно сделать вывод о том, что с вероятностью 0,95 указанное значение μ истинно.

Если применяется односторонняя проверка гипотезы, например, предполагаем, что срок батарей в действительности ниже стандарта, изменяется значение $t_{кр} = -1{,}833$. Поэтому такой результат приводит к тому, что $t = -2 < -1{,}833 = t_{кр}$, т. е. лежит вне области принятия гипотезы.

Следовательно, при уровне значимости 0,05 гипотеза отвергается, т. е. предположение о том, что срок батарей занижен, неверно.

3.5. Компьютерные технологии вычисления оценок и проверки гипотез

Пакет **Анализ данных** не содержит прямых средств (специальных программ) вычисления точечных оценок параметров генеральной совокупности типа разности двух средних $\mu_1 - \mu_2 = X_1 - \bar{X}_2$, суммы всех элементов $N\mu = NX$, разности долей $\pi_1 - \pi_2 = p_1 - p_2$.

Для получения таких оценок предварительно необходимо выйти в окно **Анализ данных** и обратиться к пункту меню **Описательная статистика** и выполнить эту процедуру. В результате выполнения программы будут получены средние выборок $\bar{X}$, $\bar{X}_1$, $\bar{X}_2$, стандартные отклонения S, S_1, S_2, которые в дальнейшем, применяя соответствующие формулы и средства Excel, можно использовать в целях получения тех или иных точечных оценок.

Для получения интервальных оценок в пакете **Анализ данных** имеется статистическая функция **Доверит**. Программа, реализующая эту функцию, предусматривает вычисления предельной ошибки $Z \cdot S_{\bar{x}}$ среднего арифметического μ генеральной совокупности.

Функция имеет следующие аргументы: **Доверит** (α, σ, *n*), где α — уровень значимости, определяемый по доверительной вероятности β = 0,9; 0,95; 0,99, на основании выражения α = 1 – β; σ — стандартное отклонение генеральной совокупности; *n* — объем выборки.

По существу в качестве σ должно фигурировать $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ или $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, либо $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}$ или $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}}$.

Однако в целях упрощения предполагается, что выборка и генеральная совокупность по объему совпадают, а объем генеральной совокупности бесконечен. Поэтому в программе определяется только значение $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Таким образом, для реализации функции **Доверит** (α, σ, *n*) следует предварительно определить стандартное отклонение σ. Для этого опять-таки необходимо обратиться к пункту **Описательная статистика**.

Реализация функции **Доверит** (α, σ, *n*) осуществляется стандартно. Необходимо выделить эту функцию в перечне статистических функций, щелкнув мышью, и нажать кнопку **ОК**. В появившемся окне ввести в поля ввода требуемые значения и снова щелкнуть кнопку **ОК**. Результатом вычисления будет значение предельной ошибки.

Например, для доверительной вероятности β = 0,95, стандартного отклонения σ = 7,15 и объема выборки *n* = 1000 следует указать такие аргументы функции: **Доверит** (0,05, 7,15, 1000). В результате получим предельную ошибку Δ = 0,44. Зная среднее μ генеральной совокупности, равное, например, 19,0, получим 95%-ный доверительный интервал (19,0 – 0,44; 19,0 + 0,44) = = (18,56; 19,44). При этом вычисление границ интервала можно выполнить средствами Excel.

Для определения доверительных интервалов разности средних двух совокупностей и разности долей пакет **Анализ данных** специальных программ не предусматривает. Поэтому стандартные ошибки разности $S_{\bar{x}_1-\bar{x}_2} = \sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}}$ или $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}$ и долей $S_{p_1-p_2} = \sqrt{S^2_{p_1} + S^2_{p_2}}$, где $S_{p_1} = \sqrt{\frac{P_1(1-P_1)}{n}}$, $S_{p_2} = \sqrt{\frac{P_2(1-P_2)}{n}}$, которые требуются для определения интервальных оценок, необходимо вычислять средствами Excel.

Предварительно по соответствующим формулам следует найти $\sigma_{\bar{x}_1}$, $\sigma_{\bar{x}_2}$ или $S_{\bar{x}_1}$, $S_{\bar{x}_2}$ либо S_{p_1}, S_{p_2}. Для этого, в свою очередь, необходимо знать σ или *S*, что требует обращения к программе **Описательная статистика**.

Для двусторонней и односторонней проверок гипотез относительно средней генеральной совокупности μ в пакете **Анализ данных** предусмотрена статистическая функция **Z-тест**.

Она имеет следующие аргументы: **Z-тест** (*массив*, μ, σ), где *массив* — последовательность данных выборки; μ — предполагаемое среднее значение генеральной совокупности; σ — стандартное отклонение генеральной совокупности. При этом если σ неизвестно как аргумент, оно может быть опущено. В этом случае в программе в качестве σ используется значение, полученное по формуле $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, на основании данных выборки.

Функция **Z-тест** вычисляет вероятность $P(\mu)$ того, что предполагаемое значение μ будет больше среднего значения выборки $\bar{X}$. Вероятность того, что $\mu \le x$, как известно, равна $1 - P(\mu)$.

Для того чтобы проверить нулевую гипотезу $\mu = \bar{X}$ заданного уровня значимости α, необходимо на основании вероятности $1 - P(\mu)$ определить значение *Z*, соответствующее этой вероятности, и установить, принадлежит ли оно интервалу принятия или отклонения гипотезы. В том случае, когда *Z* принадлежит интервалам $(-\infty, Z_{кр})$, $(Z_{кр}, +\infty)$, гипотеза α значимости отвергается. В противном случае она принимается.

Переход от вероятности $1 - P(\mu)$ к значению *Z* осуществляется на основании функции **Нормстобр** (*вероятность*), где *вероятность* — это $1 - P(\mu)$.

Рассмотрим пример, взятый из [5]. Выборка 9 замеров изготовления деталей в минутах представлена следующим массивом

X = (44, 48, 50, 46, 50, 46, 47, 51, 50). Предполагается, что время изготовления деталей распределено по нормальному закону. Это означает, что можно пользоваться стандартным Z-распределением.

Учитывая, что σ неизвестно, функцию **Z-тест** запишем в таком виде **Z-тест**(C1:C9;C10), предполагая, что массив X размещен в ячейках C1:C9, а μ = 49 — в ячейке C10.

Функции **Z-тест** и **Нормстобр** реализуются стандартно. Сначала щелчком мыши выделяются в списке статистических функций. Затем нажимается кнопка **ОК**, в соответствующие поля окон функций вводятся исходные данные и снова нажимается клавиша **ОК**.

В результате выполнения программы **Z-тест** для рассматриваемого примера получаем вероятность $P(\mu) = 0{,}894549$. Функция **Нормстобр**(1 – 0,894549) возвращает значение $Z = -1{,}25$. Для уровня значимости $\alpha = 0{,}05$ $Z_{кр} = \pm 1{,}96$.

Таким образом, $Z = -1{,}25$ попадает в интервал (–1,96, ±1,96) принятия гипотезы и она не отклоняется, поэтому 49 мин с вероятностью 0,95 можно принять в качестве среднего времени изготовления деталей.

Пакет **Анализ данных** содержит процедуру, предназначенную для проверки гипотезы о средних арифметических двух выборок, n, m. Программа называется **Двухвыборочный z-тест для средних**. Она вычисляет значение Z по формуле $Z = \dfrac{\overline{X} - \overline{Y}}{\sigma_{\overline{X}-\overline{Y}}}$, где $\overline{X}$, $\overline{Y}$ — средние арифметические выборок объемом n, m; $\sigma_{\overline{X}-\overline{Y}} = \sqrt{\sigma^2_{\overline{X}} + \sigma^2_{\overline{Y}}}$, $\sigma^2_{\overline{X}} = \left(\dfrac{\sigma_X}{\sqrt{n}}\right)^2$, $\sigma^2_{\overline{Y}} = \left(\dfrac{\sigma_Y}{\sqrt{m}}\right)^2$.

Иными словами, в этом случае предполагается, что выборки и генеральная совокупность совпадают.

Инициация процедуры выполняется стандартно: ее название выделяется в списке инструментов анализа и нажимается кнопка **ОК**. В открывшемся окне вводятся интервалы значений элементов первой и второй выборок путем указания соответствующих диапазонов ячеек листа Excel.

Например, D1:D14 и E1:E9. В поле *гипотетическая средняя разность* вводится значение 0, так как проверяется нулевая гипотеза о равенстве средних. В поле *альфа* вводится значение критерия значимости α, равное, например, 0,05. В поля *дисперсии* вводятся значения дисперсий, например 5 и 7. После этого нажимается кнопка **ОК**.

В результате выполнения процедуры выводится таблица с заголовком **Двухвыборочный z-тест для средних**. В таблице по строкам указываются средние значения двух выборок $\overline{X}$, $\overline{Y}$, заданные дисперсии, объемы выборок, гипотетическая разность средних, значения Z, наблюдаемые вероятности для односторонней и двухсторонней проверки гипотезы и критические значения $Z_{кр} = 1{,}64$ — для односторонней и $Z_{кр} = 1{,}96$ — для двухсторонней проверки.

Принять или отклонить гипотезу можно по значениям вероятностей либо по сравнению Z с $Z_{кр}$. В том случае, если вероятность для двухсторонней проверок меньше $\alpha = 0{,}05$, гипотеза отвергается. Она отвергается и в том случае, если Z принадлежит интервалу $(-\infty, Z_{кр})$ либо интервалу $(Z_{кр}, +\infty)$.

К сожалению, пакет **Анализ данных** не содержит процедур для проверки нулевых гипотез о разности долей совокупностей. Такую проверку необходимо выполнять средствами Excel.

Для проверки гипотез по критерию χ^2 в пакете **Анализ данных** предусмотрена специальная функция **ХИ2тест**. Она имеет следующие аргументы: *фактические частоты*, *теоретические частоты*. Как фактические, так и теоретические частоты задаются в виде интервалов ячеек книги Excel, например (D3:D10; I3:I10). Эти данные заносятся в соответствующие поля ввода, которые появляются на экране после выделения функции в списке статистических функций и нажатия кнопки **ОК**.

Результатом выполнения функции **ХИ2тест** является значение вероятности $P(\chi^2)$, найденное по предварительно рассчитанному значению χ^2 и числу степеней свободы.

Судят о близости эмпирического и теоретического законов распределения рассматриваемых случайных величин по значению вероятности $P(\chi^2)$, полученному в результате реализации функции **ХИ2тест**. Если вероятность $P(\chi^2) > 0{,}5$, то считается, что эмпирические и теоретические распределения близки. При $0{,}2 < P(\chi^2) \le 0{,}5$ совпадение между ними удовлетворительное. В остальных случаях — недостаточное.

Определить степень согласия между эмпирическим и теоретическим распределениями можно и путем вычисления значе-

ний χ^2 и $\chi^2_{кр}$. При этом χ^2 вычисляется по формуле (3.5), а $\chi^2_{кр}$ при помощи функции **ХИ2обр** (*вероятность, степени свободы*), в которой в качестве вероятности принимается значение уровня значимости α. В том случае, когда $\chi^2 > \chi^2_{кр}$, гипотеза о совпадении распределений отвергается. В противном случае считается, что эмпирическое и теоретическое распределения близки.

Для вычисления интервальных оценок и проверки гипотез, опирающихся на *t*-распределение Стьюдента, в пакете **Анализ данных** предусмотрена функция **Стьюдраспобр.** Она имеет следующие аргументы: **Стьюдраспобр** (*вероятность, степень свободы*).

При этом аргумент *вероятность* представляет собой уровень значимости α. Поэтому, когда вычисляется 95%-ный доверительный интервал для средней генеральной совокупности по выборке и неизвестному σ, в качестве α берется значение 0,05.

Результатом реализации функции является значение *t*, которое в дальнейшем используется для определения границ интервала по формуле (3.9), в которой фигурирует $S_{\bar{x}} = \frac{S}{\sqrt{n-1}}$, определяемое по выборке средствами Excel.

Функция реализуется стандартно: выделяется в списке статистических функций, после чего нажимается кнопка **ОК**. Далее в поля ввода соответствующего окна заносятся значения α и *d* и снова нажимается кнопка **ОК**.

Проверка гипотезы о равенстве средней генеральной совокупности μ и выборки $\bar{X}$ осуществляется с применением этой же функции. Сначала средствами Excel вычисляется значение $t = \frac{\mu - \bar{X}}{S_{\bar{x}}}$, которое далее сравнивается с $t_{кр}$, найденным при помощи функции **Стьюдраспобр** для заданного критерия значимости α и числа степеней свободы *d*.

Выше была описана процедура **Двухвыборочный z-тест для средних**, которая вычисляет значение *Z* в результате сравнения двух выборок $\bar{X}$, $\bar{Y}$ при известных дисперсиях σ_X, σ_Y в предположении, что генеральные совокупности, из которых взяты выборки, распределены нормально.

Когда σ_X, σ_Y неизвестны, а объемы выборок малы для сравнения средних, в пакете **Анализ данных** предусмотрены две процедуры: **Двухвыборочный *t*-тест для средних при неизвестных равных дисперсиях** σ_X, σ_Y и **Двухвыборочный *t*-тест для средних при неизвестных разных дисперсиях** σ_X, σ_Y. Обе процедуры позволяют вычислять значения *t*-статистики, которые, как показывает практика, по величинам достаточно близки в том случае, когда близки σ_X, σ_Y. Поэтому достаточно реализовать одну процедуру, и лучше всего первую.

Использование процедуры осуществляется стандартно: ее название выделяется в списке инструментов анализа и нажимается кнопка **ОК**. В открывшемся окне вводятся интервалы значений элементов первой и второй выборок, а также значение уровня значимости α. После этого нажимается кнопка **ОК**. Полученное значение *t* сравнивается с $t_{кр}$, на основании чего делается вывод о принятии либо непринятии гипотезы.

Рассмотрим, например, выборочные данные о расходе сырья при производстве продукции по старой и новой технологиям, которые приведены на листе Microsoft Excel по адресам D18:D26, E18:E30 [5]. При уровне значимости $\alpha = 0{,}05$ требуется проверить нулевую гипотезу, равны ли генеральные средние μ_X, μ_Y в предположении, что их дисперсии одинаковы.

В поля ввода *интервал переменной* вводятся адреса ячеек D18:D26 и E18:E30. В поле *гипотетическая разность* вводится значение 0, так проверяется отличие разности. В поле *альфа* вводится значение 0,05, после чего нажимается кнопка **ОК**.

В результате выполнения процедуры получаем таблицу, в которой для переменных *x* и *y* указываются средние арифметические, дисперсии, объемы выборок, гипотетическая разность, число степеней свободы *df*, *t*-статистика, наблюдаемые значимости $P(t)$, $t_{кр}$ для односторонней и двухсторонней проверок гипотезы.

Судить о принятии и отклонении гипотезы можно по двум показателям: наблюдаемой значимости и *t*-статистике. В данном случае наблюдаемая значимость $P(t) < 0{,}005$, а принятый уровень значимости $\alpha = 0{,}05$. Так как $P(t) < \alpha = 0{,}05$, гипотеза отклоняется. При этом $t_{кр} = 2{,}09$, а *t*-статистика равна 3,86. Вследствие того что $t = 3{,}36 > 2{,}09 = t_{кр}$, по этому соотношению гипотеза тоже отклоняется.

Кроме процедур **Двухвыборочный *t*-тест с одинаковыми дисперсиями** и **Двухвыборочный *t*-тест с различными дисперсиями** в пакете **Анализ данных** имеется статистическая функция Ттест,

которая рассчитывает вероятность t-статистики, т. е. вероятность того, что две выборки, взятые из генеральной совокупности, имеют одно и то же среднее.

Функция имеет следующие аргументы: **Ттест**(*массив*1, *массив*, *хвосты*, *тип*), где *массив*1 и *массив*2 — данные выборок X, Y; *хвосты* — число хвостов распределения: 1 или 2; *тип* — значения 1, 2, 3, определяющие тип теста. Если тип = 2, то выполняется тест в предположении, что $\sigma_X^2 \neq \sigma_Y^2$.

Функция реализуется стандартно. В результате получаем наблюдаемую значимость $P(t)$, сравнивая которую с α можно принять или отклонить рассматриваемую нулевую гипотезу.

Пакет **Анализ данных** также содержит процедуру **Двухвыборочный тест для дисперсий**. Эта процедура позволяет проверять нулевую гипотезу о равенстве двух дисперсий σ_X^2 и σ_Y^2 только на основании данных двух выборок X и Y. При этом предполагается, что наблюдения выборок распределены по нормальному закону, а отклонение от нормальности на результаты выполнения теста влияет несущественно.

В том случае, когда нулевая гипотеза $\sigma_X^2 = \sigma_Y^2$ выполняется, величина $F = \frac{S_X^2}{S_Y^2}$, называемая статистикой Фишера, имеет F-распределение с числом степеней свободы $d = n_X + n_Y - 2$. Такое распределение называют распределением дисперсионного отношения. Его используют при проверке гипотезы, равно ли $\sigma_X^2 = \sigma_Y^2$, а также во многих других случаях статистического анализа.

F-распределение по существу представляет собой отношение двух χ^2-распределений с различными числами степеней свободы. Оно зависит от значений α и степеней свободы d_1 и d_2 и представляет собой семейство распределений. Для использования F-распределения созданы специальные таблицы, имеющиеся в учебниках по статистике.

Форма F-распределения зависит от степеней свободы. Распределение по существу несимметрично. Однако по мере увеличения значений степеней свободы оно становится все более симметричным.

Процедура **Двухвыборочный тест для дисперсий** выполняется так же, как и все описанные процедуры. Выделяется в списке инструментов с последующим нажатием кнопки **ОК**, в поля открывшегося окна заносятся исходные данные и снова нажимается кнопка **ОК**.

Результатом расчетов является таблица со значениями F-статистики, $F_{кр}$ и наблюдаемой значимости $P(F)$.

Родственной по своей сущности **Двухвыборочному тесту для дисперсий** является функция **Фтест**(*массив*1, *массив*2), которая на основании двух массивов рассчитывает наблюдаемую значимость $P(F)$. При этом расчет осуществляется для двухстороннего режима (для двух хвостов).

Контрольные вопросы

1. Что понимают под числовыми оценками параметров генеральной совокупности?
2. Как классифицируют оценки параметров?
3. В чем состоит принципиальное различие между точечными и интервальными оценками?
4. Каким требованиям должны удовлетворять точечные оценки?
5. Какие виды точечных оценок используются в статистическом анализе?
6. Что представляет собой доверительный интервал среднего арифметического генеральной совокупности?
7. Как понимают термин «доверительная вероятность»?
8. Какой закон распределения случайной величины положен в основу вычисления интервальных оценок и почему?
9. Запишите и объясните общую формулу вычисления интервальных оценок параметров генеральной совокупности.
10. Почему в формуле используется стандартная ошибка выборки $S_{\bar{x}}$?
11. Запишите и объясните формулу вычисления интервальной оценки разности средних двух выборок.
12. При каких условиях биномиальный закон распределения случайной величины может быть заменен нормальным законом?

13. Запишите и объясните формулы вычисления стандартной ошибки выборки для доли.
14. Запишите и объясните формулу вычисления разности долей двух генеральных совокупностей.
15. Что в статистике понимают под гипотезой и ее проверкой?
16. Почему возникает задача проверки правильности гипотез?
17. Как истолковываются нулевая и противоположная ей гипотезы?
18. Что по существу представляет собой уровень значимости?
19. Какой закон распределения случайной величины положен в основу проверки гипотез?
20. Что представляют собой области принятия и непринятия гипотезы?
21. Что такое критические пределы?
22. Чем различаются двухсторонняя и односторонняя проверки гипотезы?
23. Запишите и объясните общее выражение вычисления Z-статистики, которое используется для проверки гипотез.
24. Что понимают под $Z_{кр}$ и как оно используется в процессе проверки гипотез?
25. Запишите и объясните выражение, используемое для проверки нулевой гипотезы о равенстве двух совокупностей.
26. Можно ли проверять гипотезы для долей и при каких условиях?
27. Можно ли проверять гипотезы о равенстве распределений вероятностей двух случайных величин?
28. Что по существу представляет собой распределение χ^2 Пирсона?
29. Каковы основы его применения для проверки гипотез о равенстве частот распределения?
30. Запишите и объясните общее выражение для вычисления значения χ^2.

31. Как вычисляется $\chi^2_{кр}$?
32. Как оно применяется для проверки гипотез о близости частот двух распределений?
33. Можно ли применять χ^2 распределение для проверки гипотез о близости эмпирического и теоретического законов распределений непрерывных и дискретных величин?
34. Как в статистике толкуется понятие «число степеней свободы»?
35. Какие проблемы возникают при вычислении доверительных интервалов и проверке гипотез для малых объемов выборок?
36. Какое распределение вероятностей случайных величин лежит в основе решения задач, связанных с выборками малых объемов?
37. Представьте кривую t-распределения Стьюдента и проведите ее сравнение с кривой нормального распределения. В чем их различие?
38. В каких случаях при малых выборках для вычисления интервальных оценок и проверки гипотез используется Z-распределение и t-распределение?
39. Запишите и объясните формулы вычисления интервальных оценок средней генеральной совокупности, суммы всех ее элементов, разности двух средних.
40. Как вычисляется значение t-распределения Стьюдента, по которому определяются предельные ошибки параметров?
41. Различаются ли принципы проверки гипотез для выборок больших и малых объемов?
42. Как определяется и используется значение $t_{кр}$ для проверки гипотез при малых выборках?
43. Какие программные средства содержит пакет **Анализ данных** для вычисления предельной ошибки среднего арифметического генеральной совокупности?
44. Содержит ли пакет **Анализ данных** программные средства для вычисления доверительных интервалов разности средних двух совокупностей и разности долей?

45. Какая статистическая функция пакета **Анализ данных** используется для проверки гипотез?
46. Какое значение позволяет вычислить эта функция?
47. Какая процедура пакета **Анализ данных** используется для проверки гипотезы о разности средних арифметических двух совокупностей?
48. Какая величина является результатом выполнения этой процедуры?
49. Содержит ли пакет **Анализ данных** программные средства для проверки гипотез о разности долей двух совокупностей?
50. Какая функция предусмотрена в пакете **Анализ данных** для проверки гипотез по критерию χ^2?
51. Какие программные средства пакета **Анализ данных** предусмотрены для вычисления $\chi^2_{кр}$?
52. Какая статистическая функция пакета **Анализ данных** используется для вычисления интервальных оценок?
53. Зачем в пакете **Анализ данных** используется процедура **Двухвыборочный тест для средних**?

Задачи

1. Из 100 баз снабжения взята случайная выборка объемом 25 баз, для которых средняя стоимость запасов оказалась равной 15 000 000 руб.
   Требуется: 1) оценить среднюю стоимость запасов каждой из 100 баз; 2) суммарную стоимость запасов 25 баз и 100 баз; 3) долю выборки и долю всех 100 баз, если из 25 баз выборки 20 баз имеют определенный товар; 4) число баз из 100, имеющих указанный товар.
2. Для случайной выборки $n = 50$ бухгалтерских счетов предприятия, взятой из 3000 счетов этого же предприятия, среднее дебетовое сальдо $\overline{X}$ оказалось равным 1696 руб., а стандартное отклонение $S = 448$ руб.

   Определить: 1) 95%-ный доверительный интервал для среднего сальдо всех 3000 счетов; 2) суммарное сальдо 3000 счетов.
   Расчет провести вручную и с использованием программных средств пакета **Анализ данных**. Объясните причины различия полученных числовых результатов.
3. Для двух предприятий розничной торговли взяты случайные выборки из 50 счетов. На одном предприятии среднее сальдо счета $\overline{X}_1$ составило 1140 руб. при стандартной ошибке $S_{\overline{X}_1} = 64$ руб., а на другом среднее сальдо $\overline{X}_2 = 1728$ руб. при $S_{\overline{X}_2} = 96$ руб.
   Требуется определить 95%-ный доверительный интервал разности средних сальдо счетов двух предприятий. Задачу решить ручным способом и на компьютере.
4. Из 100 аспирантов, случайно отобранных из нескольких университетов, 60 человек оказались сыновьями бизнесменов. Для 95%-ного доверительного интервала требуется определить долю аспирантов, являющихся сыновьями бизнесменов.
5. При общем числе аспирантов 800 человек для 95%-ного доверительного интервала требуется найти число всех аспирантов, отцы которых бизнесмены.
6. В двух густонаселенных районах города проведен опрос 100 жителей по поводу их отношения к прокладке автомобильной дороги через эти районы. В одном районе предложение поддержали 60 человек, в другом — 50.
   Для 95%-ного доверительного интервала необходимо определить разность долей лиц, поддерживающих прокладку дороги.
7. Средний возраст заказчиков ателье готового платья составляет 40 лет. Для случайной выборки из 50 человек средний возраст оказался равным 37 лет, а стандартное отклонение 12 лет. Сформулируйте и проверьте гипотезу относительно средней генеральной совокупности (среднего возраста заказчиков).
8. Изготовитель теннисных мячей утверждает, что разработал мяч, который летит в среднем на 20 м дальше за 5 с, чем стандартный мяч. В результате 40 игр оказалось, что заме-

ренная средняя скорость полета мяча на 4,2 с быстрее, чем стандартная его скорость при стандартном отклонении 1,1 с. Сформулируйте и проверьте одностороннюю гипотезу относительно средней скорости полета мяча. Задачу решите при помощи компьютерного пакета **Анализ данных**.

9. На вступительных экзаменах двух университетов были случайно отобраны 45 студентов одного университета, для которых средний бал оказался равным 552 со стандартным отклонением $S = 114$, и 38 студентов другого университета со средним баллом 530 и стандартным отклонением $S_2 = 114$. При $\alpha = 0{,}05$ проверьте гипотезу о равенстве средних баллов студентов. Используйте соответствующую программу пакета **Анализ данных**.
10. При 0,05 уровне значимости проверьте гипотезу о том, что более 30 % семей РФ имеют доступ к Интернету. Случайная выборка составляет 150 семей с результатом: 38 % семей имеют доступ к Интернету.
11. При 0,05 уровне значимости проверьте двустороннюю гипотезу о том, что 70 % заказчиков некоторой фирмы по ремонту автомобилей являются мужчинами. Объем выборки — 200 заказчиков, из которых 67 % мужчины.
12. Руководство некоторого колледжа утверждает, что в среднем количество студентов в группе равно 35 человек. Для уровня значимости 0,01 и выборок количеств студентов 42, 28, 36, 47, 35, 41, 33, 30, 39, 38 подтвердите или отклоните это утверждение. Для решения задачи используйте функцию **Стьюдобр**.
13. Проверьте гипотезу 0,05 уровня значимости о разности средних арифметических двух совокупностей при следующих данных: $\bar{X}_1 = 2{,}4$, $\sigma_1 = 0{,}6$, $n_1 = 50$; $\bar{X}_2 = 2{,}1$, $\sigma_2 = 0{,}8$, $n_2 = 60$. Используйте процедуру **Двухвыборочный z-тест для средних**.
14. Проверьте гипотезу 0,05 уровня значимости о разности средних арифметических двух совокупностей при следующих арифметических данных: $\bar{X}_1 = 49{,}2$; $n_1 = 12$; $\bar{X}_2 = 41{,}1$; $n_2 = 10$. Стандартные отклонения выборок предполагаются неизвестными, но равными. Используйте процедуру **Двухвыборочный t-тест для средних**.

15. Проверьте ту же гипотезу в предположении, что стандартные отклонения выборок равные.
16. Один из футболистов некоторой команды высшей лиги заявляет, что вероятность забить им мяч в любой игре составляет 30 %. Таблица частот забитых мячей за последние полгода представлена ниже.

Таблица 3.4. **Частоты забитых мячей**

Количество мячей	Количество игр
0	26
1	34
2	30
3	7
4	3

Предполагается, что распределение частот забитых мячей биномиальное. При уровне значимости $\alpha = 0{,}05$ и $P = 0{,}3$ проверьте эту гипотезу.

Глава 4
ДИСПЕРСИОННЫЙ, КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

4.1. Дисперсионный анализ

В предшествующей главе были описаны методы проверки гипотез о равенстве средних арифметических двух совокупностей в предположении, что они независимы и распределены нормально. Рассмотрены также программные средства, предлагаемые Microsoft Excel для решения задач с известными и неизвестными дисперсиями двух совокупностей. В этих методах для вычисления наблюдаемого уровня значимости использовалось как нормальное Z-распределение, так и t-распределение Стьюдента.

Но как осуществить сравнение средних трех и более совокупностей? Один из способов — попарное сравнение, когда для каждой пары совокупностей применяются методы, рассмотренные ранее. Этот метод весьма трудоемок и очень часто требует учета того факта, что некоторые совокупности могут быть зависимыми. Поэтому в статистике для случая сравнения средних трех и более совокупностей применяют специальный метод, получивший название «Дисперсионный анализ» (*ANOVA — Analysis of Variance*).

Идеи этого метода основываются на том, что в случае отклонения нулевой гипотезы о равенстве средних двух выборок неравенство средних объясняют воздействием на случайные переменные выборок нескольких факторов. При этом конкретную реализацию фактора называют уровнем фактора $A^{(i)}$, $i = 1, 2, ..., m$, а наблюдаемое значение случайного признака, например среднего арифметического выборки, называют результативным признаком и обозначают Y.

Пусть $\overline{X}_1, \overline{X}_2, ..., \overline{X}_m$ — средние арифметические выборок, полученные в результате воздействия уровней фактора $A^{(1)}, A^{(2)}, ...$

$..., A^{(m)}$. Если при изменении уровня фактора значения $\overline{X}_1 = \overline{X}_2 = ... = \overline{X}_m$, т. е. не изменяются, считают, что результативный признак не зависит от фактора. В противном случае полагают, что такая связь существует.

В зависимости от числа факторов, оказывающих воздействие на результативный признак, различают однофакторный и многофакторный дисперсионный анализ. При этом как однофакторный, так и многофакторный анализ позволяет получить правдоподобные статистические заключения в том случае, когда выборки независимы, каждая из выборок имеет нормальное распределение и дисперсии выборок одинаковы. Методика дисперсионного анализа состоит в сравнении дисперсий (разбросов) средних.

Рассмотрим пример. Президент торговой компании требует выяснить, существует ли различие между средним числом покупателей четырех магазинов и чем оно объясняется. Данные, полученные в результате подсчета покупателей, посетивших магазины в течение дня на протяжении недели, приведены в табл. 4.1.

Таблица 4.1. **Число покупателей магазинов**

Магазин 1	Магазин 2	Магазин 3	Магазин 4
36	35	26	26
48	20	20	52
32	31	38	37
28	22	32	36
31	19	37	18
55	42	15	30
29	0	21	0
$\overline{X}_1 = 37{,}0$	$\overline{X}_2 = 28{,}2$	$\overline{X}_3 = 27{,}0$	$\overline{X}_4 = 33{,}2$
$S_1^2 = 108{,}5$	$S_2^2 = 86{,}96$	$S_3^2 = 79{,}34$	$S_4^2 = 133{,}76$

При этом средние значения $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$, $\overline{X}_4$ и дисперсии S_1^2, S_2^2, S_3^2, S_4^2 вычислены известным способом по выражениям (1.1), (1.9), а полученные значения средних показали, что они различны для разных магазинов.

Необходимо установить, случаен ли разброс средних, или он закономерен и вызван определенной причиной? Таким образом, в этом примере фактором является число покупателей магазинов и рассматривается четыре уровня действия этого фактора — посещение покупателями четырех магазинов.

Для того чтобы проверить, случаен ли разброс среднего числа покупателей магазинов или не случаен, необходимо проверить нулевую гипотезу: $\bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \bar{X}_4$? Если средние не равны, гипотеза отклоняется и мы приходим к заключению, что разброс средних значений не случаен. В противном случае гипотеза подтверждается, т. е. разброс случаен.

В свою очередь, проверка гипотезы основывается на сравнении разброса (дисперсии) между средними выборок с разбросом (дисперсией) внутри выборок. Или, как принято говорить, межвыборочных и внутривыборочных дисперсий.

Пусть $i = 1, 2, \ldots, r$ — число выборок; n_i, $\bar{x}_i$ — объем и среднее арифметическое i-й выборки. Для вычисления дисперсий между выборками используется взвешенная по объемам выборок сумма квадратов отклонений групповых средних от общей выборочной средней:

$$S_m = \sum_{i=1}^{r} n_i (\bar{x}_i - \bar{\bar{x}})^2. \qquad (4.1)$$

При этом $\bar{\bar{x}}$ вычисляется по выражению

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{r} \bar{x}_i}{r}.$$

Применительно к рассматриваемому примеру получаем:

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{x}_4}{4} = \frac{37 + 28{,}2 + 27 + 33{,}2}{4} = 31{,}4.$$

На основании этого

$$S_m = 7(37 - 31{,}4)^2 + 6(28{,}2 - 31{,}4)^2 + 7(27 - 31{,}4)^2 + 6(33{,}2 - 31{,}4)^2 =$$

$$= 219{,}52 + 61{,}44 + 135{,}52 + 19{,}44 = 435{,}92.$$

Внутривыборочная дисперсия определяется по выражению

$$S_v = \sum_{i=1}^{r} (n_i - 1) S_i^2. \qquad (4.2)$$

В нашем случае она равна:

$$(7 - 1)108{,}5 + (6 - 1)86{,}96 + (7 - 1)79{,}34 + (6 - 1)133{,}76 =$$

$$= 651 + 434{,}8 + 476{,}04 + 668{,}8 = 2230{,}64.$$

Между собой сравнивается несмещенная межвыборочная дисперсия, определяемая по выражению $\bar{S}_m = \frac{S_m}{r - 1}$, и несмещенная внутривыборочная дисперсия $\bar{S}_v = \frac{S_v}{\sum_{i=1}^{r} n_i - r}$.

На основании этих формул для рассматриваемого примера получаем:

$$\bar{S}_m = \frac{435{,}92}{3} = 145{,}31 \quad \text{и} \quad \bar{S}_v = \frac{2230{,}64}{26 - 4} = 101{,}39.$$

В математической статистике доказано, что отношение $F = \frac{\bar{S}_m}{\bar{S}_v}$ подчиняется F-распределению Фишера со степенями свободы $d_1 = r - 1$ и $d_2 = \sum_{i=1}^{r} n_i - r$. Поэтому проверка гипотезы о равенстве средних $\bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \bar{X}_4$ сводится к вычислению значения F и сравнению его с $F_{кр}$, которое вычисляется для принятого уровня значимости α и заданных степеней свободы d_1 и d_2.

В нашем примере $F = \frac{145{,}31}{101{,}39} = 1{,}43$, а $F_{кр}$ при $\alpha = 0{,}05$ и $d_1 = 3$, $d_2 = 23$ равно 3,028. В связи с тем что $F = 1{,}42 < 3{,}028 = F_{кр}$, гипотеза о равенстве средних подтверждается. Таким образом, влияние фактора — разное посещение покупателями четырех магазинов — выборочными исследованиями не подтвердилось. На этом основании президент компании может считать, что число посе-

тителей магазинов за определенный период с вероятностью 0,95 одинаково.

Рассмотрим еще один пример из [5]. Прораб некоторой стройки интересуется вопросом, зависит ли объем выполненной работы на некотором объекте от работающей бригады.

Выборочные данные объемов работы, выполненных четырьмя бригадами за четыре смены, приведены в табл. 4.2.

Таблица 4.2. **Объем работ, выполненных бригадами**

Номер смены	Бригада 1	Бригада 2	Бригада 3	Бригада 4
1	140	150	148	150
2	144	149	149	155
3	142	152	146	154
4	145	150	147	152
Среднее	$\bar{X}_1 = 142{,}75$	$\bar{X}_2 = 150{,}25$	$\bar{X}_3 = 147{,}5$	$\bar{X}_4 = 152{,}75$
Дисперсия	$S_1^2 = 4{,}92$	$S_2^2 = 1{,}58$	$S_3^2 = 1{,}67$	$S_4^2 = 4{,}92$

Для уровня значимости $\alpha = 0{,}05$ проверим гипотезу: равны ли средние объемы работ, т. е. $\bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \bar{X}_4$? Для этого по выражениям (4.1), (4.2) вычислим межвыборочную и внутривыборочную дисперсии.

Имеем:

$$\bar{\bar{X}} = \frac{142{,}75 + 150{,}25 + 147{,}5 + 152{,}75}{4} = 148{,}3.$$

Далее:

$$S_m = 4(142{,}75 - 148{,}3)^2 + 4(150 - 148{,}3)^2 +$$
$$+ 4(147{,}5 - 148{,}3)^2 + 4(152{,}75 - 148{,}3)^2 =$$
$$= 123{,}20 + 11{,}56 + 2{,}56 + 79{,}20 = 216{,}52.$$

Для внутривыборочной дисперсии получаем:

$$S_v = 3\ (4.92 + 1{,}58 + 1{,}67 + 4{,}92) = 39{,}27.$$

Несмещенная межвыборочная дисперсия $\bar{S}_m = \frac{216{,}52}{3} = 72{,}17.$

Несмещенная внутривыборочная дисперсия $\bar{S}_v = \frac{39{,}27}{12} = 3{,}27.$

Таким образом: $F = \frac{72{,}17}{3{,}27} = 22{,}07.$

Для $\alpha = 0{,}05$ и числа степеней свободы $d_1 = 3$, $d_2 = 12$ получаем $F_{кр} = 3{,}49$. Так как $F = 22{,}07 > 3{,}49 = F_{кр}$, гипотеза о равенстве средних объемов работ, выполняемых каждой бригадой, отвергается. На основании этого с вероятностью 0,95 можно считать, что объем выполненных работ зависит от работающей бригады.

Однако дисперсионный анализ не позволяет сравнивать средние совокупностей между собой и определять, какое из них больше остальных. Эта задача решается при помощи метода Шеффе.

Проверка Шеффе предусматривает сравнение каждой пары выборок. В нашем случае необходимо сравнить $\bar{X}_1$ и $\bar{X}_2$, $\bar{X}_1$ и $\bar{X}_3$, $\bar{X}_1$ и $\bar{X}_4$, $\bar{X}_2$ и $\bar{X}_3$, $\bar{X}_2$ и $\bar{X}_4$, $\bar{X}_3$ и $\bar{X}_4$. Процедуре сравнения предшествует вычисление для каждой пары критерия Шеффе по выражению

$$F_{Sh} = \frac{(\bar{X}_i - \bar{X}_{i+1})^2}{S_v \Big/ \sum_{i=1}^{r}(n_i - 1)\left(\frac{1}{n_i} + \frac{1}{n_{i+1}}\right)}, \qquad (4.3)$$

где $\bar{X}_i$, $\bar{X}_{i+1}$ — средние i-й, i+1-й выборки;

n_i, n_{i+1} — объемы i-й и i+1-й выборок.

На основании выражения (4.3) получаем:

$$F_{Sh}^{1,2} = \frac{(\bar{x}_1 - \bar{x}_2)^2}{12\left(\frac{1}{4} + \frac{1}{4}\right)} = \frac{(142{,}75 - 150{,}25)^2}{6} = \frac{56{,}25}{6} = 9{,}37;$$

$$F_{Sh}^{1,3} = \frac{(\bar{x}_1 - \bar{x}_3)^2}{6} = \frac{(142{,}75 - 147{,}5)^2}{6} = \frac{22{,}75}{6} = 3{,}76;$$

$$F_{Sh}^{1,4} = \frac{(\bar{x}_1 - \bar{x}_4)^2}{6} = \frac{(142{,}75 - 152{,}75)^2}{6} = \frac{100}{6} = 16{,}67;$$

$$F_{Sh}^{2,3} = \frac{(\bar{x}_2 - \bar{x}_3)^2}{6} = \frac{(150{,}25 - 147{,}5)^2}{6} = \frac{7{,}5625}{6} = 1{,}26;$$

$$F_{Sh}^{2,4} = \frac{(\bar{x}_2 - \bar{x}_4)^2}{6} = \frac{(150{,}25 - 150{,}75)^2}{6} = \frac{0{,}25}{6} = 0{,}042;$$

$$F_{Sh}^{3,4} = \frac{(\bar{x}_3 - \bar{x}_4)^2}{6} = \frac{(147{,}5 - 152{,}75)^2}{6} = \frac{27{,}5625}{6} = 4{,}59;$$

$$F_{Sh}^{2,4} = \frac{(\bar{x}_2 - \bar{x}_4)^2}{6} = \frac{(150{,}25 - 150{,}75)^2}{6} = \frac{0{,}25}{6} = 0{,}042.$$

Критическое значение критерия Шеффе вычисляется так:

$$F_{Sh}^{кр} = (r - 1)F_{кр}. \qquad (4.4)$$

В результате получаем $F_{Sh}^{кр} = (4 - 1)3{,}49 = 3 \cdot 3{,}49 = 10{,}47$.

Если $F_{Sh} \le F_{Sh}^{кр}$, различие между средними отсутствует. Сравнивая вычисленные значения F_{Sh} с $F_{Sh}^{кр}$, получаем:

$F_{Sh}^{1,2} = 9{,}37 < 10{,}47 = F_{Sh}^{кр}$; $F_{Sh}^{1,3} = 3{,}76 < 10{,}47 = F_{Sh}^{кр}$;

$F_{Sh}^{1,4} = 16{,}67 > 10{,}47 = F_{Sh}^{кр}$; $F_{Sh}^{2,3} = 1{,}26 < 10{,}47 = F_{Sh}^{кр}$;

$F_{Sh}^{2,4} = 0{,}042 < 10{,}47 = F_{Sh}^{кр}$; $F_{Sh}^{3,4} = 4{,}59 < 10{,}47 = F_{Sh}^{кр}$.

Таким образом, в соответствии с полученными результатами сравнений статистическое различие в выработке имеется между первой и четвертой бригадами.

В том случае, когда на результирующий признак воздействуют два фактора A, B, возникает задача двухфакторного дисперсионного анализа. Логика однофакторного и двухфакторного дисперсионного анализа во многом схожа, и с правилами ее компьютерной реализации можно ознакомиться в [5].

4.2. Корреляционный анализ

Весьма часто в финансово-экономической деятельности возникает задача установления уровня связи между изучаемыми переменными. Например, отдел рекламы предприятия может интересовать, как увеличивается объем продаж от качества рекламы.

Ректор университета желает знать, есть ли связь между объемом часов, отводимых на определенную дисциплину, и экзаменационной оценкой студентов. Покупателей электронной техники может заинтересовать вопрос, как зависит цена телевизора от размера экрана. Очевидно, что число подобных примеров легко умножить.

Статистика располагает специальными методами, которые более-менее правдоподобно позволяют ответить на эти вопросы. Это методы корреляционного и регрессионного анализа.

Оба названия латинского происхождения и означают взаимозависимость и возвращение назад. При этом корреляционный анализ завершается установлением тесноты и направления связи между переменными, а регрессионный анализ — построением эмпирической зависимости, связывающей переменные. Практически оба вида исследований применяют для того, чтобы предсказать поведение одних величин на основании изменения других.

При определении связи между переменными прежде всего определяют, какая переменная является возможной причиной, а какая — следствием связи.

В приведенных примерах причина — это качество рекламы, объем часов, отводимых на предмет, и размер экрана телевизора. Следствие — объем продаж, экзаменационная оценка и цена телевизора. Однако связь не во всех случаях может быть причиной.

Например, в некоторой семье увеличиваются доходы, вследствие чего растут расходы и сбережения. Главу семьи может заинтересовать, как зависят расходы и сбережения от доходов. Очевидно, что связь между величинами носит причинный характер. В то же время можно поинтересоваться и тем, существует ли связь между сбережениями и затратами, хотя явная причина такой связи не наблюдается.

Поэтому корреляционный анализ предполагает изучение как причинных, так и воображаемых связей между переменными.

Для того чтобы абстрагироваться от конкретных объектов и формализовать будущие действия, в статистике, как и в математике, при определении взаимосвязи между переменными вводят понятие независимых и зависимых переменных. Таким образом, независимые переменные — это причины или предполагаемые причины, зависимые — следствия причинной или предполагаемой связи.

В математике связь между независимой и зависимой переменными определяется функцией. Говорят, что если каждому элементу x множества $X = \{x\}$ соответствует один и только один элемент y множества $Y = \{y\}$, то на множестве X определена функция $y = f(x)$. По существу функция — это правило, согласно которому каждый x отображается в y.

В статистике, где изучаются случайные события и величины, именно из-за случайности указать такое правило не всегда представляется возможным. Для одного значения независимой переменной оно может быть одно, для другой — другое и т. д. Поэтому при определении взаимосвязи между переменными оперируют понятием не функциональной, а вероятностной зависимости, т. е. зависимости не детерминированной, а возможной.

Весьма наглядным инструментом изучения взаимосвязи (корреляции) являются *диаграммы рассеяния*. На рис. 4.1 представлены различные виды этих диаграмм в виде точек (x, y) случайных величин x, y.

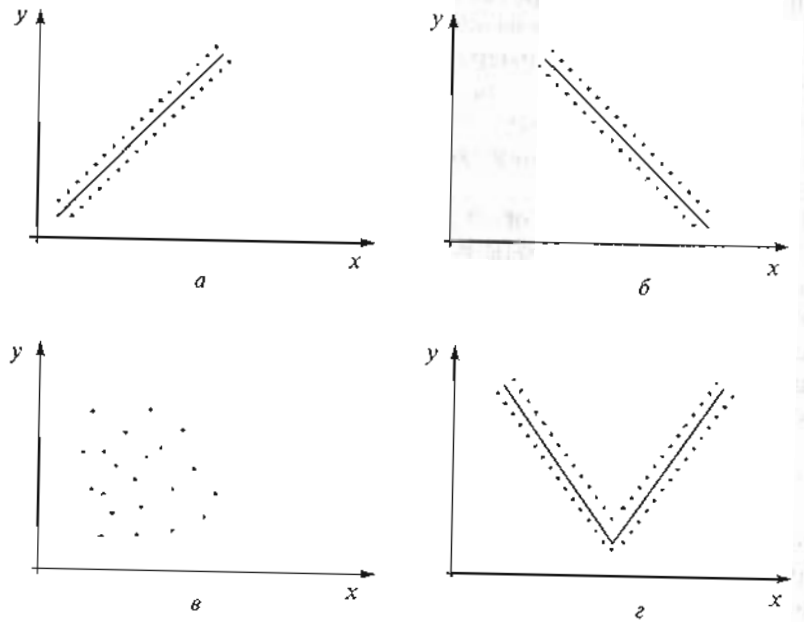

Рис. 4.1. Диаграммы рассеяния случайных величин x, y:
а — положительная линейная корреляция; *б* — отрицательная линейная корреляция; *в* — отсутствие корреляции; *г* — нелинейная корреляция

Построение диаграмм рассеяния является первым шагом корреляционного анализа, поскольку наглядно представляет вид связи между независимой и зависимой переменными.

Так, на диаграмме *а* рис. 4.1 изображена положительная линейная связь, поскольку с увеличением переменной x возрастает y. Причем этот рост примерно линеен, что следует из прямой линии, проведенной относительно этих точек, уравнение которой $y = ax + b$.

На диаграмме *б* показана отрицательная линейная корреляция, поскольку с увеличением x уменьшается y, причем так же, как и для диаграммы *а*, линейно. Вид линейной функции для этого случая такой: $y = -ax + b$.

Диаграмма *в* показывает отсутствие связи между переменными, а на диаграмме *г* изображена нелинейная связь, так как вначале с ростом x значение y уменьшается, а затем возрастает.

Как и в математике, рассматривающей функции многих переменных, в статистике изучаются связи не только множества переменных (x, y), а и пар (X, Y), где X — фиксированные наборы переменных (векторы). Такие связи называются многофакторными, в связи с чем корреляционный анализ делится на однофакторный и многофакторный (часто говорят, парный и множественный анализ).

Для измерения тесноты линейной связи в парной корреляции применяют коэффициент корреляции К. Пирсона. Он вычисляется по различным формулам, наиболее часто употребляемая из которых приведена ниже:

$$r_{xy} = \frac{n\sum_{i=1}^{n}(x_i y_i) - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}, \qquad (4.5)$$

где (x_1, y_1), (x_2, y_2), ..., (x_i, y_i), ..., (x_n, y_n), $i = 1, 2, \ldots, n$ — пары переменных, из которых x_i — независимая, а y_i — зависимая переменная.

Эта формула используется для вычисления коэффициента корреляции r_{xy} как для генеральной совокупности, т. е. когда n исчерпывает все элементы совокупности, так и для выборки, ко-

где n — объем выборки, представляющий часть элементов генеральной совокупности. При этом коэффициент корреляции r_{xy} характеризует только линейную вероятностную связь переменных, которые должны быть непрерывными. В тех случаях, когда эта связь нелинейная, для измерения тесноты используют корреляционное отношение, часто называемое индекс корреляции [6].

В том случае, когда случайные величины x, y связаны точной линейной зависимостью $y = ax + b$ или $y = -ax + b$, коэффициент корреляции $r_{xy} = \pm 1$. Когда же связь не функциональная, а вероятностная, коэффициент корреляции изменяется в пределах $-1 < r_{xy} < +1$. Если $0 \le r_{xy} < 1$, связь считается прямой (положительной), в противном случае, т. е. когда $-1 < r_{xy} \le 0$, — обратной (отрицательной).

Например, если расходы семьи на образование детей по мере увеличения уровня семейного дохода возрастают, коэффициент корреляции $0 \le r_{xy} < 1$ отражает прямую связь между этими величинами. Когда же при возрастании личной задолженности объем планируемых закупок снижается, коэффициент корреляции $-1 < r_x \le 0$ отражает обратную связь. Значение коэффициента $r_{xy} = 0$ указывает на отсутствие какой-либо связи между переменными.

На практике степень связи часто определяют по шкале Чеддока, представленной в табл. 4.3.

Таблица 4.3. **Шкала Чеддока тесноты корреляционной связи**

Теснота связи	Коэффициент корреляции, r_{xy}	
	Прямая связь	Обратная связь
Слабая	0,1 – 0,3	(–0,1)–(–0,3)
Умеренная	0,3–0,5	(–0,3)–(–0,5)
Заметная	0,5–0,7	(–0,5)–(–0,7)
Высокая	0,7–0,9	(–0,7)–(–0,9)
Весьма высокая	0,9–0,99	(–0,9)–(–0,99)

Вместе с тем полученный на основании выборок коэффициент корреляции r_{xy}, характеризует только эти выборки, и для того, чтобы сделать выводы о коэффициенте корреляции генеральной

совокупности, необходимо провести его полный статистический анализ. Для этого требуется определить интервальную оценку этого коэффициента и проверить нулевую гипотезу о его значимости. Указанные действия выполняются в несколько этапов.

Вначале проверяется нулевая гипотеза, равен ли $r_{xy} = 0$. Доказано, что величина $t = \sqrt{\dfrac{r_{xy}^2 (n-2)}{1 - r_{xy}^2}}$, где n — объем выборки, имеет t-распределение Стьюдента, с $(n-2)$ степенями свободы. Поэтому, если $t > t_{кр}$ для заданного уровня значимости α и $(n-2)$ степеней свободы гипотеза отвергается, т. е. коэффициент r_{xy} значим и он отражает взаимосвязь случайных x, y генеральной совокупности. В противном случае такой связи нет.

Если коэффициент значим, вычисление доверительного интервала предусматривает преобразование этого коэффициента в Z-преобразование нормально распределенной случайной величины. Для этого служит формула

$$Z = \frac{1}{2} \ln \frac{1 + r_{xy}}{1 - r_{xy}}, \qquad (4.6)$$

которая носит название прямого Z-преобразования Фишера.

После вычисления значения Z границы доверительного интервала для Z определяются по выражениям

$$\left(Z - Z_{кр} \sqrt{\frac{1}{n-3}} \right); \left(Z + Z_{кр} \sqrt{\frac{1}{n-3}} \right),$$

где $Z_{кр}$ — критическое значение Z для уровня доверия $1 - \alpha$.

После того как они найдены, выполняется обратное преобразование левой и правой границ в значения r_{xy} коэффициента корреляции. Для этого используется формула $r_{xy} = \dfrac{e^{2z} - 1}{e^{2z} + 1}$, которая носит название обратного преобразования Фишера.

Стандартное отклонение коэффициента корреляции рассчитывается по выражению

$$S_r = \sqrt{\frac{1 - r_{xy}^2}{1 + r_{xy}^2}}.$$

Приведенные формулы выполнения статистического анализа коэффициента линейной корреляции для ручного счета весьма трудоемки. Поэтому в пакете **Анализ данных** предусмотрены специальные статистические функции, позволяющие в стандартном режиме Microsoft Excel существенно ускорить этот процесс. Указанные функции будут рассмотрены в соответствующем параграфе этой главы.

Методы множественного или многофакторного корреляционного анализа применяются в тех случаях, когда необходимо измерить тесноту связи между двумя или более независимыми переменными, с одной стороны, и одной зависимой переменной — с другой. Например, если необходимо установить тесноту связи между уровнем задолженности некоторого лица по кредиту, доходом его семьи и текущим банковским счетом.

Для вычисления степени корреляции следует применить метод множественного корреляционного анализа. В этом случае независимыми переменными являются пары: доход семьи, текущий банковский счет. Зависимая переменная — уровень задолженности по кредиту.

Таким образом, в методе парной корреляции имеют дело с двумя векторами переменных $X = (x_1, x_2, x_i, \ldots, x_n)$, $Y = (y_1, y_2, y_i, \ldots \ldots, y_n)$.

В методах множественной корреляции приходится оперировать с вектором $Y = (y_1, y_2, \ldots, y_i, \ldots, y_n)$ и матрицей — прямоугольной таблицей значений независимых переменных

$$X_m = \begin{bmatrix} x_{11}, & x_{12}, & \ldots, & x_{1i}, & \ldots, & x_{1n} \\ x_{21}, & x_{22}, & \ldots, & x_{2i}, & \ldots, & x_{2n} \\ x_{m1}, & x_{m2}, & \ldots, & x_{mi}, & \ldots, & x_{mn} \end{bmatrix},$$ где m — число независимых переменных, используемых в анализе.

Вычисления, проводимые для оценки тесноты связи для модели множественной корреляции, весьма трудоемки и сводятся к получению матрицы коэффициентов парной корреляции с дальнейшим их статистическим анализом коэффициента множественной корреляции r_m, получаемого на основании этой матрицы.

В заключение отметим, что независимо от того, коэффициент какой корреляции вычисляется — парной или множественной, он определяет степень зависимости между переменными, но эта зависимость не обязательно должна быть причинной. Причинность или ее отсутствие определяется на уровне постановки той или иной задачи.

4.3. Построение эмпирических формул

Эмпирические формулы — это аналитические выражения, связывающие зависимую (зависимые) и независимую (независимые) переменную (переменные), полученные на основании эксперимента. Они показывают, как численно определяется зависимая переменная y, если задано некоторое значение независимой переменной x, и определяют характер изменения зависимой переменной при изменении независимой.

Эмпирические формулы нужны в основном для того, чтобы предсказать поведение зависимой вне диапазона изменения независимой переменной. Это выполняется методами, имеющими название «электрополяция функций».

Эмпирические формулы получают на основании регрессионного анализа, и они отражают связь между переменными, полученными на основании опыта или наблюдений, что и определяет название этих формул.

Собственно говоря, регрессионный анализ — это раздел статистики, посвященный получению и статистическому обоснованию эмпирических зависимостей между случайными переменными.

В свою очередь, такое название этот анализ получил от прямых линий наилучшего положения, которые английский статистик Ф. Гальтон привел, изучая зависимость между ростом родителей и их детей. Он обнаружил: высокие родители имеют меньших по росту детей, а низкие — более высоких.

Такая зависимость, по мнению Ф. Гальтона, свидетельствовала о регрессии (возврате) к среднему в наследственных чертах. Поэтому полученные линии, отражающие связь между указанными величинами, он назвал линиями регрессии, опираясь на свой вывод о том, что природа не терпит крайностей.

Различают линейный парный регрессионный анализ и линейный множественный анализ. И тот, и другой вид анализа подразумевает, что зависимые и независимые переменные связаны между собой линейной зависимостью типа $y = ax + b$. При

этом в парном анализе оперируют двумя векторами $X = (x_1, x_2, ..., x_i, ..., x_n)$, $Y = (y_1, y_2, ..., y_i, ..., y_n)$, а в множественном анализе вместо вектора X используется матрица X_m вида

$$X_m = \begin{bmatrix} x_{11}, & x_{12}, & \ldots, & x_{1i}, & \ldots, & x_{1n} \\ x_{21}, & x_{22}, & \ldots, & x_{2i}, & \ldots, & x_{2n} \\ x_{m1}, & x_{m2}, & \ldots, & x_{mi}, & \ldots, & x_{mn} \end{bmatrix},$$

в которой m — число факторов, воздействующих на зависимую переменную Y.

Кроме этих двух видов регрессионного анализа существует метод парного нелинейного анализа и множественного нелинейного анализа. В них предполагается нелинейная зависимость между переменными, для которых определяется линия регрессии.

Парный линейный анализ, как и корреляционный, обычно начинают с построения диаграммы рассеяния в плоскости xOy. Это объясняется тем, что по расположению точек (x, y) на плоскости xOy визуально можно определить, линейная или нелинейная зависимость между переменными.

Задача парного линейного регрессионного анализа состоит в том, чтобы для заданного набора эмпирических точек (x, y) провести линию наилучшего положения типа $\hat{y} = ax + b$. Такой линией статистики считают ту линию, для которой сумма квадратов отклонений от экспериментальных точек (x, y) минимальна.

На рис. 4.2 представлена диаграмма рассеяния, линия $\hat{y} = ax + b$ и показано, сумму каких величин требуется минимизировать.

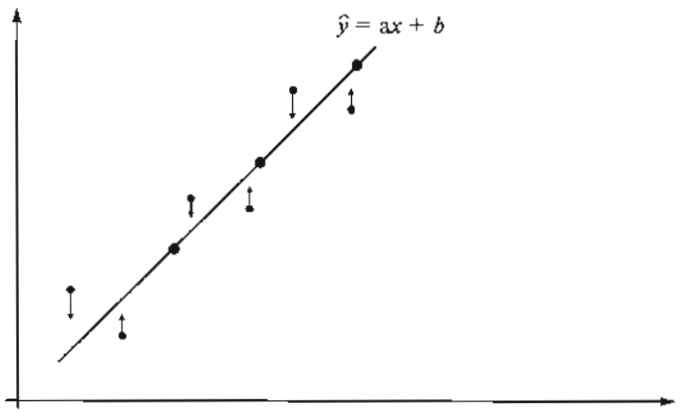

Рис. 4.2. К минимизации суммы квадратов отклонений

Величины, составляющие сумму, изображены вертикальными векторами, связывающими эмпирические точки (x, y) с точками линии $\hat{y} = ax + b$.

По существу векторы, в принятом масштабе, представляют собой отклонения эмпирических точек (x, y) от точек линии.

Таким образом, задачу поиска линии, удовлетворяющей этому требованию, можно записать в виде $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min$.

Способ ее решения получил название «метод наименьших квадратов». Он позволяет найти параметры a, b линии $\hat{y} = ax + b$, т. е., по существу, определить эту линию так, что будет выполняться основное требование — достигаться минимум суммы квадратов отклонений.

Для вычисления параметров a, b получены следующие выражения:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}; \quad b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \qquad (4.7)$$

При этом, когда вычислено значение a, для получения b можно использовать выражение $b = \bar{Y} - a\bar{X}$, где $\bar{Y}$, $\bar{X}$ — средние значения для y_i, x_i, $i = 1, 2, ..., n$.

Рассмотрим пример. Заданы эмпирические данные x_i, y_i, $i = 1, 2, ..., n$, представленные в табл. 4.4. (колонки 1, 2). В колонках 3 и 4 показаны этапы расчета коэффициента a.

Подставим полученные значения в выражение для вычисления a. В результате получим:

$$a = \frac{10 \cdot 658 - 55 \cdot 105}{385 - 55^2} = 0{,}976.$$

Далее: $b = 10{,}5 - 0{,}976 \cdot 5{,}5 = 5{,}13$.

Таким образом, уравнение линии регрессии будет иметь такой вид: $\hat{y} = 0{,}976 \cdot x + 5{,}13$. Оно указывает на то, что с увеличением значения x зависимая переменная $\hat{y}$ будет также возрастать. При этом параметр $a = 0{,}976$ определяет тангенс угла наклона прямой к оси Ox и тем самым задает степень роста y при

Таблица 4.4. Расчет коэффициента a уравнения $\hat{y} = ax + b$

x	y	xy	x^2
1	8	8	1
2	6	12	4
3	10	30	9
4	6	24	16
5	10	50	25
6	13	78	36
7	9	63	49
8	11	88	64
9	15	135	81
10	17	170	100
$\sum_{i=1}^{10} x_i = 55$	$\sum_{i=1}^{10} y_i = 105$	$\sum_{i=1}^{10} x_i y_i = 658$	$\sum_{i=1}^{10} x_i^2 = 385$

увеличении x, а параметр $b = 5{,}13$ определяет величину y, отсекаемую прямой $\hat{y} = 0{,}976 \cdot x + 5{,}13$ на оси Oy при $x = 0$.

Уравнение линии регрессии $\hat{y} = ax + b$, полученное методом наименьших квадратов, часто называемое теоретическим уравнением, лишь приближенно оценивает набор эмпирических значений y_i, $i = 1, 2, \ldots, n$. В том случае, когда эмпирическая точка (x_i, y_i) лежит на прямой $\hat{y} = ax + b$, ошибка приближения равна нулю, когда же это условие нарушается, она отлична от нуля.

В связи с тем что набор y_i, $i = 1, 2, \ldots, n$ — это набор случайных величин, ошибка приближения (аппроксимации) эмпирических точек теоретической прямой может быть охарактеризована разбросом этих точек относительно прямой $\hat{y}$, т. е. стандартным отклонением S_{xy} от этой прямой. Стандартное отклонение S_{xy} для этого случая вычисляется по выражению

$$S_{xy} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2}}, \qquad (4.8)$$

где $\hat{y}_i$ — значения y, лежащие на прямой $\hat{y} = ax + b$.

Знаменатель $(n - 2)$ в выражении (4.7) определяется числом степеней свободы, равным числу пар (x_i, y_i) за вычетом двух фиксированных значений, определяемых коэффициентами a, b уравнения $\hat{y} = ax + b$.

Подкоренное выражение в формуле стандартного отклонения S_{xy} представляет собой дисперсию D_{xy}, т. е. разброс эмпирических значений y_i, $i = 1, 2, \ldots, n$ относительно точек прямой $\hat{y}$, $i = 1, 2, \ldots, n$. В отличие от общей дисперсии $D_y = \frac{\sum_{i=1}^{n} (y_i - \bar{Y})^2}{n}$, представляющей разброс эмпирических значений y_i, $i = 1, 2, \ldots, n$ относительно среднего значения $\bar{Y}$, дисперсия D_{xy} называется остаточной.

Сравнение этой дисперсии с общей дисперсией показывает, насколько меньше разброс эмпирических значений y_i, $i = 1, 2, \ldots, n$, относительно точек прямой $\hat{y}$, чем их разброс относительно средней $\bar{Y}$. Тем самым это сравнение характеризует качество приближения прямой $\hat{y} = ax + b$ к набору точек (x_i, y_i), $i = 1, 2, \ldots, n$.

В целом уравнение $\hat{y} = ax + b$ рассматривают как случайный объект, а для того, чтобы правомерно его использовать, т. е. распространить выводы на генеральную совокупность, уравнение подвергают общему статистическому анализу: проверяют нулевые гипотезы о значимости уравнения $\hat{y}$ и его коэффициентов a, b, а также определяют доверительные их интервалы.

Статистическую значимость уравнения $\hat{y} = ax + b$ оценивают на основании сравнения общей D_y и остаточной D_{xy} дисперсий методом Фишера. Для этого при степенях свободы $(n - 1)$ и $(n - 2)$ и заданном уровне значимости α по F-распределению Фишера определяют критическое значение $F_{кр}$.

Если оказывается, что $F = \frac{D_y}{D_{xy}} < F_{кр}$, то с вероятностью $1 - \alpha$ уравнение $\hat{y} = ax + b$ адекватно представляет данные y_i генеральной совокупности. В противном случае это не так.

На практике, в частности в компьютерных технологиях регрессионного анализа, применяют другой способ проверки адекватности уравнения $\hat{y} = ax + b$ эмпирическим данным.

Он состоит в том, что вместо соотношения между общей и остаточной дисперсиями используют так называемый *коэффициент детерминации*, определяемый по выражению

$$R_d^2 = \frac{D_F}{D_y},$$

где $D_F = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2}{n}$.

Тогда значение F определяется по формуле

$$F = \frac{R_d^2 \cdot (n-1)}{1 - R_d^2}.$$

Доказано, что F имеет F-распределение Фишера с одной и $(n-1)$ степенями свободы. Это дает возможность при заданном уровне значимости α и заданном n определить $F_{кр}$. Если окажется, что $F > F_{кр}$, нулевая гипотеза о том, что $R_d^2 = 0$ отвергается, т. е. уравнение $\hat{y} = ax + b$ считается адекватным эмпирическим данным генеральной совокупности.

Доверительный интервал $\hat{y}$ для любого значения x_i, $i = 1, 2, ..., n$ определяется на основании соотношения

$$\hat{y} \pm t_{кр} \cdot S_{xy}, \qquad (4.8),$$

где $t_{кр}$ — критическое значение t-распределения Стьюдента, найденное для принятого уровня доверия $1 - \alpha$ и числа степеней свободы $(n-2)$.

Гипотеза о значимости коэффициентов a, b линии $\hat{y} = ax + b$ проверяется также с использованием t-распределения Стьюдента. Для этого вычисляется значение $t = \dfrac{a}{S_a}$, где S_a — стандартное отклонение коэффициента a, определяемое по выражению

$$S_a = \frac{S_{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}}.$$

В том случае, когда $t_{кр}$, найденное для заданного уровня значимости α и числа степеней свободы $(n-2)$, окажется меньше $t = \dfrac{a}{S_a}$, т. е. $|t| > |t_{кр}|$, гипотеза α-значимости о том, что $a = 0$, отвергается и для a можно найти доверительный интервал по выражению $a \pm t_{кр} \cdot S_a$.

В связи с тем что a представляет собой тангенс угла наклона линии регрессии к оси Ox, проверка гипотезы о значении $a = 0$, по существу является проверкой того, действительно ли прямая $\hat{y} = ax + b$ не горизонтальна. Вычисление доверительного интервала для параметра a говорит о том, в каких пределах изменяется угол наклона прямой $\hat{y} = ax + b$ к оси Ox.

Практика статистического анализа показывает, что наряду с линейной зависимостью между наборами переменных y_i, x_i, $i = 1, 2, ..., n$, встречаются и другие зависимости: степенная $y = ax^b$, обратная $y = \dfrac{1}{ax + b}$, экспоненциальная $y = ae^x$ и различные модификации этих зависимостей.

Метод наименьших квадратов путем линеаризации этих зависимостей (замены переменных) позволяет подбирать соответствующие кривые, отображающие указанные взаимосвязи. Безусловно, эмпирические формулы можно подобрать путем ручного счета, остановившись на той форме кривой, которая дает наименьшую остаточную дисперсию. Однако многие специализированные пакеты программ статистической обработки данных, в частности пакет Statgraphics, содержит компьютерные средства, значительно облегчающие эту работу.

При изучении различных экономических данных весьма часто возникают ситуации, когда изучаемая случайная величина зависит не от одного, а от множества факторов. Например, экономиста может интересовать зависимость прибыли предприятий от стоимости основных фондов и величины оборотных средств.

Чтобы построить эмпирическую формулу такой связи и осуществить ее статистический анализ, используют методы множественного линейного и нелинейного регрессионного анализа.

Уравнение множественной линейной регрессии имеет обычно такой вид:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + ... + a_mx_m,$$

где $x_1, x_2, ..., x_m$ — величины факторов;

$a_0, a_1, ..., a_m$ — коэффициенты уравнения.

Как и в случае парной линейной регрессии, коэффициенты этого уравнения определяют методом наименьших квадратов. Далее проводят статистическую оценку качества полученного эмпирического уравнения, после чего определяют статистическую значимость каждого коэффициента уравнения и вычисляют их доверительные интервалы.

Все перечисленные действия осуществляют методами, которые были описаны при рассмотрении статистического анализа парной линейной регрессии.

Методы нелинейного регрессионного анализа сводятся к тому, что уравнение регрессии представляют в виде полинома такой степени $\hat{y} = a_0 + a_1 x_1 + a_2 x_2^2 + \ldots + a_e x_e^e$, при которой достигается минимальная остаточная дисперсия. Анализ такой зависимости достаточно трудоемок и выходит за рамки данного учебного пособия. Рекомендуемая литература по этому вопросу [7].

4.4. Компьютерные технологии дисперсионного, корреляционного и регрессионного анализа

Пакет **Анализ данных** Microsoft Excel для проведения однофакторного дисперсионного анализа представляет процедуру **Однофакторный дисперсионный анализ**.

Вызов процедуры стандартный: в списке инструментов анализа необходимо выделить указанную процедуру и щелкнуть кнопку **ОК**. В результате на экране появится окно с полями ввода, в которые необходимо ввести исходные данные, ссылаясь на диапазон ячеек книги Excel, куда они предварительно должны быть внесены. После этого следует снова нажать кнопку **ОК**.

Результаты выполнения процедуры выводятся на экран в виде двух таблиц. Одна содержит объемы выборок, значения среднего $\bar{X}$ и дисперсии выборок, другая — межгрупповую, внутригрупповую и общую дисперсии, обозначенные *SS*; числа степеней свободы, обозначенные *df*; несмещенные дисперсии, обозначенные *MS*; наблюдаемую значимость $P(F)$; значение F-статистики и $F_{кр}$.

На основании сопоставления F-статистики с $F_{кр}$ или наблюдаемой значимости $P(F)$ со значимостью $P(\alpha)$ принимается или

отвергается гипотеза о равенстве средних арифметических факторов.

Необходимо отметить, что исходные данные, помещаемые в поле процедуры, представляются одним диапазоном ячеек. Например, выработки каждой из четырех бригад в каждую смену (см. табл. 4.2) следует ввести одним диапазоном A1:D4 или B1:E4.

Процедура, реализующая метод Шеффе, в пакете **Анализ данных** не представлена. В связи с этим расчет согласно ее методике следует выполнять стандартными средствами Microsoft Excel.

Компьютерные технологии корреляционного анализа реализуются процедурой **Корреляция** и рядом статистических функций.

Процедура **Корреляция** вызывается и выполняется стандартно. Результат ее выполнения — коэффициент линейной корреляции r_{xy}. Такой же результат может быть получен и при помощи функции **Коррел** (*массив*1, *массив*2), где *массив*1, *массив*2 — массивы независимой и зависимой переменных, представленные диапазонами ячеек листа Excel. Функция выполняется стандартно: выделяется в списке статистических функций, вызывается окно ввода (нажимается кнопка **ОК**), вводятся исходные данные и снова нажимается кнопка **ОК**.

В качестве примера использования процедуры **Корреляция** и функции **Коррел** рассмотрим вычисление коэффициентов корреляции для данных, характеризующих уровень образования — независимая переменная *x* и уровень преступности — зависимая переменная *y* — для ряда областей России [5]. При этом уровень образования определялся как число жителей со средним и высшим образованием, приходящихся на 1000 человек, а уровень преступности — как число преступлений на 100 000 жителей (табл. 4.5).

Таблица 4.5. **Исходные данные к расчету коэффициента корреляции**

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x	735	788	779	795	740	902	838	763	762	757	772	764	755	755
y	908	791	804	701	685	496	536	936	662	671	920	1040	882	882

Исходные данные для процедуры **Корреляция** представим диапазоном ячеек A1:B14 и поместим их в поле *входной интервал*. В результате выполнения процедуры получим значение ко-

эффициента $r_{xy} = -0{,}66$. Он свидетельствует о том, что связь между переменной x и y заметная и отрицательная, т. е. с повышением уровня образования уровень преступности падает.

То же значение коэффициента корреляции получим и в результате применения функции **Коррел**.

Кроме функции **Коррел** для выполнения статистического анализа коэффициента корреляции r_{xy} предусмотрены функции **Фишер** (r_{xy}) и **Фишер** (Z).

Применение этих функций объясним, следуя порядку статистического анализа коэффициента r_{xy}, изложенному в п. 4.2 настоящей главы.

1. При помощи функции **Коррел** вычислим значение r_{xy}. Как уже говорилось, он равен 0,66.

2. По выражению $t = \sqrt{\dfrac{r_{xy}^2 (n-2)}{1 - r_{xy}^2}}$ средствами Excel определим значение t-статистики для r_{xy}. Оно равно 3,043.

3. При помощи функции **Стьюдраспобр** (α, d) для принятого уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $d = 14 - 2 = 12$ определим $t_{кр}$. В результате этого получим $t_{кр} = 2{,}1788$.

4. Проверим гипотезу о равенстве $r_{xy} = 0$. Для этого сравним t-статистику с $t_{кр}$. Так как $t = 3{,}043 > 2{,}1788 = t_{кр}$, гипотеза о равенстве $r_{xy} = 0$ отвергается, что свидетельствует о значимости коэффициента корреляции, а следовательно, о статистической зависимости между переменными x, y.

5. Теперь найдем интервальную оценку для коэффициента r_{xy}. Для этого при помощи функции **Фишер** (r_{xy}) найдем Z-преобразование коэффициента корреляции r_{xy}. В результате получим $Z = -0{,}7928$.

6. Доверительный интервал для Z, как было указано в п. 4.2, определяется так: $\left(Z - Z_{кр}\sqrt{\dfrac{1}{n-3}},\ Z + Z_{кр}\sqrt{\dfrac{1}{n-3}} \right)$.

В результате для 95%-ного доверительного интервала Z получаем: $\left(-0{,}7928 - 1{,}96\sqrt{\dfrac{1}{14-3}},\ -0{,}7928 + 1{,}96\sqrt{\dfrac{1}{14-3}} \right)$. Следовательно, левая граница равна $-1{,}384$, а правая — 0,208.

7. Используя функцию **Фишеробр**, найдем левую и правую границы интервала изменения коэффициента корреляции $r_{xy} = -0{,}66$. Получаем: левая граница равна $-0{,}88$, правая — 0,205.

8. По формуле $S_r = \sqrt{\dfrac{1 - r_{xy}^2}{n-2}}$ найдем стандартную ошибку линейного коэффициента корреляции. Она равна 0,217.

Методы регрессионного анализа, предусмотренные пакетом **Анализ данных**, включают процедуру **Регрессия** и целый ряд статистических функций.

Процедура **Регрессия** реализует метод множественной линейной регрессии так, что теоретическое уравнение представляется в виде $\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m$, где a_0, a_1, ..., a_m — коэффициенты уравнения, а x_1, x_2, ..., x_m — значения факторов.

Вызов процедуры и ее выполнение осуществляется стандартным способом. При этом в процессе заполнения полей ввода необходимо придерживаться следующих правил.

Входные интервалы множеств Y и X должны содержать ссылки на ячейки книги Excel, содержащие данные по зависимым и независимым переменным. Данные по Y представляются одним столбцом, данные по X — не более чем 16 столбцами.

В поле *уровень надежности* заносится значение уровня надежности $1 - \alpha$, используемое при анализе коэффициента детерминации R_d^2 и коэффициентов уравнения a_0, a_1, ..., a_m. Если $\alpha = 0{,}05$, т. е. уровень доверия (надежности) 95 %, это поле пропускается.

Поле *константа-ноль*, если оно активно, предусматривает прохождение линии регрессии через начало координат, т. е. $a_0 = 0$.

Поля *остатки* и *стандартизированные остатки*, если они активны, предусматривают вывод остатков, т. е. разностей $\hat{y}_i - y_i$, $i = 1, 2, \ldots, n$.

Поля *график остатков*, *график подбора*, *график нормальной вероятности*, в активном их состоянии, обеспечивают вывод на экран соответственно графика, остатков зависимости $\hat{y} = f(x)$ и зависимости y от интервалов персентилей.

После выполнения процедуры на выходе получаем таблицу регрессивной статистики, включающую следующие данные: коэффициент множественной корреляции — r_m, коэффициент де-

терминации — R_d^2, нормированный коэффициент — R_{dn}^2, стандартную ошибку $S_{xy} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ и объем выборки.

Далее приводится таблица дисперсионного анализа с числами степеней свободы *df* для количества факторов, остаточной и полной дисперсии, со значениями факторной, остаточной и полной дисперсии *SS*, со значениями несмещенной дисперсии *MS*, значением *F*-статистики Фишера и наблюдаемой *F*-значимости.

Кроме этого, выводится таблица коэффициентов уравнения a_1, a_1, ..., a_m и их статистического анализа, включающего стандартные ошибки, *t*-статистики Стьюдента, наблюдаемые *P*(*F*)-значимости, доверительные интервалы.

Безусловно, процедура **Регрессия** может успешно применяться и для анализа парной линейной регрессии. В этом случае набор независимых переменных x_i, $i = 1, 2, ..., n$ представляется одним столбцом книги Excel. Кроме того, на экран можно вывести график линии регрессии $\hat{y} = ax + b$.

О том, насколько хорошо отражает линия регрессии линейную связь между множеством зависимых переменных *X* и зависимых *Y*, обычно судят по коэффициенту детерминации R_d^2. Если он близок к единице, связь считается идеальной, т. е. линия регрессии лучшим образом отображает эту связь.

Для данных табл. 4.5 в результате выполнения процедуры **Регрессия** получено значение коэффициента корреляции $r_{xy} = -0{,}66$, детерминаци $R_d^2 = 0{,}43$, стандартной ошибки $S_{xy} = 123{,}3$. Эти данные показывают, что линейная зависимость $\hat{y} = ax + b$ не совсем удачно отображает действительную связь между множествами переменных *X*, *Y*.

Данные, получаемые в результате выполнения процедуры **Регрессия**, можно получить также при помощи статистической функции **Линейн**(*Y*, *X*, *конст*, *статистика*), аргументы которой *Y*, *X* — соответственно множества зависимых и независимых переменных; *конст* = 1 или 0, или опущен; *статистика* = 1 или 0, или опущен.

В том случае, когда аргумент *конст* = 0, коэффициент уравнения $a_0 = 0$, т. е. линия регрессии проходит через начало координат. Когда аргумент *статистика* = 1, выводятся результаты статистического анализа уравнения $\hat{y} = a_0 + a_1x_1 + ... + a_nx_n$.

Функцию **Линейн** удобно применять в том случае, когда не требуется проводить статистического анализа уравнения регрессии. Безусловно, она может быть использована для построения уравнения парной линейной регрессии.

Аргумент *X* функции **Линейн** необязателен. Если он опущен, то вводятся значения *x*, равные 1, 2, 3, ..., под которыми подразумеваются либо номера 1, 2, 3, ..., массивов размера *Y*, либо в случае парной регрессии значения независимой переменной 1, 2, 3.

При помощи этой функции для данных табл. 4.5 получены значения $a = -2{,}376$, $b = 2629$.

Для экстраполяции функции $\hat{y} = f(x)$, т. е. вычисления ее значений за пределами значений зависимой переменной, для которой определялась линия регрессии, в пакете **Анализ данных** предусмотрена статистическая функция **Тенденция**(*Y*, *X*, $X_{нов}$, *конст*), аргументы которой $X_{нов}$ — те значения независимой переменной, для которых по уравнению $\hat{y} = f(x)$ требуется получить значения зависимой переменной. Аргументы *X*, $X_{нов}$ могут быть опущены. Тогда предполагается, что они равны.

Если аргумент *конст* = 0, линия регрессии проходит через начало координат. Например, для $x = 1000$ значение *y*, полученное применением этой функции, равно 250,2.

Частным случаем функции **Тенденция**(*Y*, *X*, $X_{нов}$, *конст*) является функция **Предсказ**(*X*, *Y*, X_p), в которой *X* — варианты *x*, для которых необходимо вычислить (предсказать) значение $\hat{y}$.

Статистические функции **Наклон**(*Y*, *X*) и **Отрезок**(*Y*, *X*) вычисляют угол наклона *a* линии $\hat{y} = ax + b$ и значение *b* для парной линейной регрессии.

Полученные при помощи этих функций значения $a = -2{,}376$, $b = 2629$. Это совпадает с данными, найденными ранее при помощи функции **Линейн**.

Функция **Стош**(*Y*, *X*) для парной линейной регрессии определяет стандартное отклонение S_{xy}. Для данных табл. 4.5 оно равно 123,25, что совпадает с результатом, полученным процедурой **Регрессия**.

Пакет **Анализ данных** предусматривает возможность построения линии регрессии, отображающей экспоненциальную (показательную) зависимость между множествами переменных $X = \{x\}$, $Y = \{y\}$ вида $\hat{y} = a_0 \cdot a_1^{x_1} \cdot a_2^{x_2} \cdot ... \cdot a_m^{x_m}$, где x_1, ..., x_m — значения факторов, а a_0, a_1, a_2, ..., a_m — коэффициенты уравнения.

Этой цели служит статистическая функция **Лгрфприбл**(*Y*, *X*, *конст*, *статистика*). Если ее аргумент *конст* = 0, то коэффициент $a_0 \neq 1$. Если аргумент *статистика* = 1, выводятся результаты статистического анализа уравнения $\hat{y} = a_0 \cdot a_1^{x_1} \cdot a_2^{x_2} \cdot \ldots \cdot a_m^{x_m}$. В противном случае, т. е. когда аргумент *статистика* = 0, выводятся только значения коэффициентов a_0, a_1, a_2, ..., a_m.

Указанная функция позволяет строить линии регрессии как для множественной, так и для парной регрессии.

Экстраполировать экспоненциальную зависимость позволяет функция **Рост**(*Y*, *X*, $X_{нов}$, *конст.*), где $X_{нов}$ — новые значения факторов *X*, для которых необходимо получить дополнительные значения *Y*; *конст.* — аргумент, определяющий $a_0 = 1$ или 0, по аналогии с функцией **Лгрфприбл**.

Контрольные вопросы

1. Какой метод статистического анализа применяется для сравнения более двух средних арифметических совокупностей случайных данных?
2. Какие идеи положены в основу этого метода?
3. В чем смысл понятий «фактор» и «результативный признак»?
4. В чем заключается различие однофакторного и многофакторного дисперсионного анализа?
5. На сравнении каких величин основывается проверка нулевой гипотезы о равенстве средних арифметических нескольких совокупностей в методе дисперсионного анализа?
6. По какому выражению определяется межвыборочная дисперсия?
7. Какое выражение служит для определения внутривыборочной дисперсии?
8. Какое распределение вероятности применяется при оценке значимости гипотезы о равенстве средних нескольких совокупностей в дисперсионном анализе?
9. Какой вывод делают в том случае, когда гипотеза подтверждается или не подтверждается?

10. Позволяют ли результаты выполненного дисперсионного анализа статистически сравнить средние совокупностей между собой по величине?
11. Как можно осуществить такое сравнение?
12. В чем смысл метода Шеффе и какие выводы можно сделать после его реализации?
13. Какие программные средства содержит пакет **Анализ данных** для решения задач дисперсионного анализа?
14. Зачем в статистике применяют корреляционный анализ и какими показателями оцениваются его результаты?
15. В чем смысл регрессионного анализа? Зачем он применяется?
16. Какие связи между случайными переменными рассматриваются в статистике?
17. С каких действий обычно начинают проводить корреляционный анализ?
18. По какой шкале определяется теснота связи?
19. Нужен ли в статистическом анализе коэффициент корреляции, если требуется обобщить тесноту связи на генеральную совокупность?
20. Какие действия включает статистический анализ коэффициента корреляции?
21. Чем отличается парный корреляционный анализ от множественного?
22. Какая компьютерная процедура из пакета **Анализ данных** предназначена для определения коэффициента корреляции?
23. Для чего предназначена статистическая функция **Коррел**?
24. Каковы назначения функций **Фишер** и **Фишеробр**?
25. Из каких действий состоит процесс статистического анализа коэффициента корреляции?
26. Что представляет собой эмпирическая формула? Чем она отличается от теоретической?
27. Зачем нужны эмпирические формулы?

28. Каким методом получают эмпирические формулы?
29. Чем различаются парный и множественный регрессионный анализ?
30. В чем смысл метода наименьших квадратов?
31. Какие величины определяются в результате применения метода наименьших квадратов к исходным данным в случае линейной регрессии?
32. Точно ли отражает уравнение линейной регрессии $\hat{y} = ax + b$ набор эмпирических значений зависимой переменной?
33. Какая величина служит для оценки качества приближения прямой $\hat{y} = ax + b$ набора случайных значений зависимой переменной?
34. Нужно ли проводить статистический анализ качества уравнения $\hat{y} = ax + b$?
35. Как оценивают статистическую значимость уравнения $\hat{y} = ax + b$?
36. Как определяется доверительный интервал для значений $\hat{y}_i$, $i = 1, 2, \ldots \ldots, n$?
37. Как определяются доверительные интервалы для коэффициентов a, b уравнения $\hat{y} = ax + b$?
38. Какой вид имеет уравнение множественной линейной регрессии? Отличается ли методика его статистического анализа от методики анализа парной линейной регрессии?
39. Какие нелинейные связи между переменными можно проанализировать при помощи линейного анализа?
40. Какая компьютерная процедура регрессионного анализа содержится в пакете **Анализ данных** Microsoft Excel?
41. Какие результаты выполнения этой процедуры выводятся на экран?
42. Для каких целей предназначена статистическая функция **Линейн**?
43. Чем различаются функции **Тенденция** и **Предсказ**? Каково их назначение?

44. Для чего предназначены статистические функции **Наклон** и **Отрезок**?
45. Можно ли с помощью функции **Стош** определить стандартное отклонение S_{xy}?
46. Какая функция позволяет построить нелинейную регрессионную зависимость типа экспоненты?
47. Для чего предназначена функция **Рост**?

Задачи

1. Имеется три участка земли, на которых выращивается трава для футбольного поля. На участки вносятся разные удобрения, и когда трава вырастает, ее скашивают и взвешивают. Требуется выяснить, влияет ли тип удобрений на рост травы, и если это так, то какой тип удобрений предпочтительнее? Данные, полученные в результате шести скашиваний для каждого участка, представлены в табл. 4.6.

Таблица 4.6. **Веса скошенной травы (кг)**

№ п/п	Участки земли		
	Удобрение 1	Удобрение 2	Удобрение 3
1	10,2	11,6	8,1
2	8,5	12,0	9,0
3	8,4	9,2	10,7
4	10,5	10,3	9,1
5	9,0	10,3	9,1
6	8,1	12,5	9,5

Прежде чем решать основную задачу, определите, что является фактором, сколько его уровней, что является выборкой и каков ее объем.
Сформулируйте нулевую и альтернативную гипотезы. Для проверки нулевой гипотезы принятого уровня значимости используйте процедуру **Однофакторный дисперсионный**

анализ пакета **Анализ данных**. Для выявления предпочтительного типа удобрения проведите анализ Шеффе.

2. Эксперты проверяют средний расход бензина трех различных моделей автомобилей с одинаковым объемом цилиндров. Для этого несколько машин каждой модели сделали пробег по 500 км, в результате чего был зафиксирован расход бензина, представленный в табл. 4.7.

Таблица 4.7. **Расход бензина автомобилями (л)**

№ п/п	Расход бензина		
	Модель 1	Модель 2	Модель 3
1	22,5	18,7	17,2
2	20,8	19,8	18,0
3	22,0	20,4	21,1
4	23,6	18,0	19,8
5	21,3	21,4	18,6
6	22,5	19,7	18,9

Требуется установить, равны ли в среднем по расходу бензина проверяемые модели машин? Если они не равны, то какая модель предпочтительнее?

Определите, что является фактором в данной задаче и сколько уровней фактора, что является выборкой, каков ее объем.

Сформулируйте нулевую и альтернативную гипотезы. Для проверки нулевой гипотезы используйте процедуру **Однофакторный дисперсионный анализ**.

Для выявления того, какая модель автомобиля предпочтительнее, проведите анализ Шеффе.

3. В табл. 4.8 приведены данные о часах, отводимых учебной частью для изучения предмета «Статистика» в течение семестра, и результатах экзаменов (в баллах), полученных студентами по этой дисциплине.

Таблица 4.8. **Исходные данные для корреляционного анализа**

Часы	30	50	40	40	20	30
Результаты экзаменов	86	95	92	83	78	82

Необходимо установить, есть ли корреляция между количеством часов, отводимых на изучение предмета, и экзаменационной оценкой студентов. Если связь между величинами существует, определите коэффициент корреляции и проведите его статистический анализ.

Коэффициент корреляции определите при помощи процедуры **Корреляция** и функции **Коррел**. Для статистического анализа используйте функции **Фишер**, **Стьюдраспобр**, **Фишеробр**. По шкале Чеддока установите степень корреляции между переменными.

4. В табл. 4.9 приведены данные об оплате (в млн долл.) десяти команд высшей баскетбольной лиги за 2008 г. с указанием числа побед каждой команды.

Таблица 4.9. **Оплата команд и число побед**

Команда	1	2	3	4	5
Оплата	$171 \cdot 10^6$	$108 \cdot 10^6$		$43 \cdot 10^6$	$58 \cdot 10^6$
Победы	103	75	92	55	56
Команда	6	7	8	9	10
Оплата	$56 \cdot 10^6$	$62 \cdot 10^6$	$43 \cdot 10^6$	$57 \cdot 10^6$	$75 \cdot 10^6$
Победы	62	84	78	73	67

Требуется установить, существует ли корреляция между оплатой команды и числом побед. Если она существует, определите ее степень, используя шкалу Чеддока. Проведите статистический анализ коэффициента корреляции.

5. В табл. 4.10 представлены данные о пробеге восьми моделей автомобилей и их цены.

Таблица 4.10. **Пробег, км, и цена автомобилей, руб.**

Пробег	21 800	34 000	41 700	53 560	65 800	72 100	76 500	84 700
Цена	16 000	11 500	13 400	14 800	10 500	12 300	8200	9500

6. Проведите парный регрессионный анализ данных, используя процедуру **Регрессия**. Выведите на экран дисплея график линии регрессии.
7. Проведите парный регрессионный анализ данных, используя статистическую функцию **Линейн**.
8. При помощи функции **Линейн** определите только коэффициенты уравнения регрессии.
9. Для значений пробега 90 000 и 100 000 определите цены автомобилей, используя статистические функции **Тенденция** и **Предсказ.**
10. При помощи функций **Наклон** и **Отрезок** определите значения коэффициентов *a*, *b* линии регрессии.
11. При помощи функции **Стош** определите стандартную ошибку линии регрессии.
12. В табл. 4.11 представлена зависимость прибыли предприятий от величины оборотных средств и стоимости основных фондов [5].

Таблица 4.11. **Исходные данные**

Номер предприятия	Прибыль, млн руб.	Величина оборотных средств, млн руб.	Стоимость основных фондов, млн руб.
1	352	115	510
2	72	59	190
3	86	69	230
4	310	87	470
5	52	42	110
6	161	135	445

13. Графическим путем определите зависимости: прибыль — величина оборотных средств, прибыль — стоимость основных фондов.
14. Если указанные зависимости линейные, проведите парный регрессионный анализ, используя процедуру **Регрессия** или функцию **Линейн**.
15. Если зависимости нелинейные, используя функцию **Лгрфприбл**, проведите регрессионный анализ этих зависимостей.
16. Используя статистическую функцию **Лгрфприбл**, проведите множественный регрессионный анализ зависимости прибыли от величины оборотных средств и стоимости основных фондов.
17. Используя функцию **Рост**, определите, какими оборотными средствами и основными фондами должно располагать предприятие, чтобы получить прибыль 400 и 500 млн руб.

Глава 5
ВРЕМЕННЫЕ РЯДЫ И ИНДЕКСЫ

5.1. Анализ временны́х рядов

Временной ряд — это последовательность случайных величин или событий, представленных через определенные, обычно равные, промежутки времени. Часто говорят последовательность данных, развернутых во времени.

Приведем примеры временны́х рядов, которые, безусловно, легко умножить: ежегодное количество продаж легковых автомобилей в России начиная с 1999 по 2008 г., т. е. в течение последних десяти лет; ежедневное количество покупателей супермаркета в течение полугода с 1 января 2008 по 30 июня 2008 г.; ежегодная добыча золота в России в течение последних десяти лет, т. е. с 1 января 1999 по 31 декабря 2008 г.; ежемесячная зарплата металлургов за последние пять лет, т. е. за последние шестьдесят месяцев.

Анализ временных рядов, во-первых, позволяет обнаружить и установить закономерности развития общественных и природных явлений за тот или иной промежуток времени, во-вторых, составить прогноз их развития на определенную перспективу.

В рядах выделяют два основных параметра: уровень, т. е. значение случайной величины в каждый момент времени, и собственно время. С позиций математики ряд — это функция дискретного типа, поскольку задана в отдельных точках временно́й оси $t_1, t_2, \ldots, t_m$.

Различают ряды интервальные (периодические) и моментные. Все приведенные примеры рядов представляют собой интервальные ряды. Моментные ряды определяют показатели на заданный момент времени. Обычно это ряды, описывающие запасы или остатки. Например, запасы бензина в бензоколонке на конец декабря, остаток денежных средств клиентов некоторого

банка к концу года, наличные оборотные фонды некоторого предприятия на начало планового периода и т. п.

По способу выражения уровней ряды делятся на ряды абсолютных, средних и относительных величин.

В табл. 5.1 показаны все перечисленные типы рядов [6].

Таблица 5.1. **Результаты строительства жилплощади в РФ**

Названия	Годы				
	1980	1985	1992	1995	2000
Число квартир, тыс.	1190	1151	682	602	373
Средний размер жилплощади, м^2	49,9	54,4	60,8	68,2	81,1
Удельный вес однокомнатных квартир, %	18,0	18,0	18,0	18,0	20,0

Анализ временного ряда, в какой бы форме ни были представлены его уровни, состоит в выявлении основных факторов, влияющих на эти уровни. В качестве таких факторов выделяют тенденцию, сезонные, циклические и неопределенные факторы.

Весьма часто анализ ряда начинают с графического его представления. На рис. 5.1 приведен пример такого представления, где изображен предполагаемый график временного ряда объема продаж некоторых видов продуктов магазинами торговой сети.

Рис. 5.1. График ряда и его компоненты

Наиболее важным фактором, влияющим на уровни ряда, принято считать тенденцию (от англ. *trend*).

Тенденция является долгосрочной составляющей ряда, определяющей общее увеличение или уменьшение его уровней во времени. На рис. 5.1 тенденция отображена прямой линией, обозначенной буквой *T*.

Сезонная компонента ряда характеризует изменение уровней, которые периодически повторяются через равные промежутки времени. Причем это изменение должно завершаться в пределах года и повторяться из года в год, чтобы классифицироваться как сезонная компонента.

Например, рост объема продаж удобрений каждую весну и сокращение его в течение остальных месяцев года может служить примером сезонного фактора временного ряда. На графике эта компонента отображена волнистой линией, обозначенной буквой *S*.

В дополнение к тенденции и сезонной компоненте в рядах выделяют так называемую циклическую компоненту. Циклические изменения похожи на сезонные, однако отличаются большей длительностью изменения. На рис. 5.1 циклическая компонента отмечена символом *C*.

Неопределенные (нерегулярные) компоненты, которые также характерны для временных рядов, представляют собой быстрые изменения, как правило, малой длительности. Например, ежедневные и еженедельные колебания продажи удобрений, которые связаны с изменениями погоды, способствуют образованию некоторых изменений малой длительности, составляющих нерегулярную компоненту временного ряда.

На рис. 5.1 эта компонента отображена мелкими колебаниями и обозначена символом *I*.

Классический метод анализа временных рядов состоит в установлении воздействия на уровни временного ряда каждой из перечисленных компонент по очереди. Этот метод получил название *разложения временного ряда*. При этом исходной является одна из двух моделей временного ряда: аддитивная или мультипликативная.

Когда в основу полагается аддитивная модель, считают, что на уровень ряда воздействует сумма компонент $T + S + C + I$, откуда и происходит ее название (от англ. *add* — складывать). Если же в основе анализа ряда полагают мультипликативную модель, то считают, что на уровни ряда воздействует произведение компонент $T \cdot S \cdot C \cdot I$.

На практике чаще всего при анализе рядов используется мультипликативная модель.

Обычно анализ рядов начинают с выделения тренда, т. е. линии, которая определяет основную долгосрочную закономерность изменения уровней ряда. Затем определяют влияние сезонной компоненты и завершают анализ установлением воздействия циклической и нерегулярной компонент.

При определении тренда можно использовать простейший ручной метод, когда линия тренда на графике ряда наносится «на глаз». По этой линии в дальнейшем определяется прогноз тенденции ряда за его временными пределами.

Более совершенными методами аналитического характера являются метод полусредних, скользящих полусредних, экспоненциального сглаживания и наименьших квадратов.

Метод полусредних состоит в том, что временной интервал разбивается на две равные или примерно равные части, для каждой из которых определяется средний уровень ряда. Таким образом, получают две точки (y_1, t_1), (y_2, t_2), через которые проводят прямую линию, определяемую как тренд.

На рис. 5.2 приведен график ряда регистрации новых автомобилей в миллионах штук в некоторой стране за 15 лет начиная с 1994 по 2008 г. На графике показаны линии тренда, нанесен-

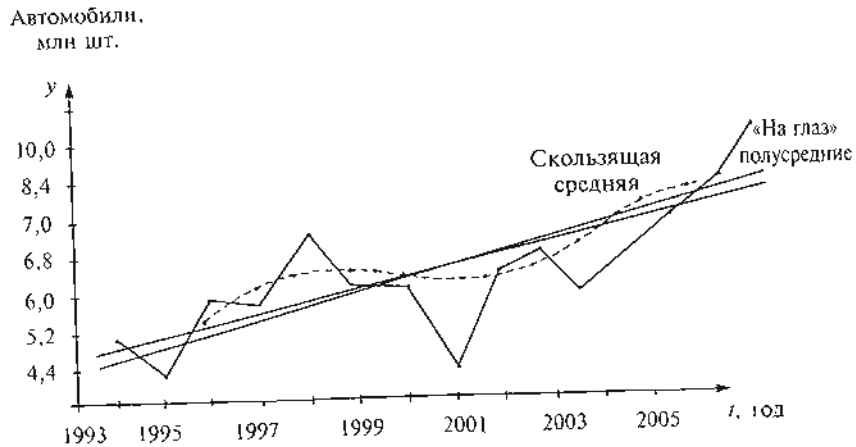

Рис. 5.2. График регистрации автомобилей

ные «на глаз», а также полученные методом полусредних. Исходные данные приведены в колонках 1, 2 табл. 5.2.

Таблица 5.2. **Временной ряд регистрации автомобилей**

Год	Зарегистрированные автомобили, млн шт.	Пятилетняя скользящая сумма	Пятилетняя скользящая средняя	Экспоненциальное сглаживание, $\alpha = 0{,}7$
1994	5,061			
1995	4,158			5,061
1996	5,739	27,633	5,533	4,429
1997	5,535	28,557	5,711	5,346
1998	7,170	30,381	6,076	6,623
1999	5,995	29,297	5,859	6,182
2000	5,982	29,803	5,961	6,042
2001	4,655	29,210	5,842	5,072
2002	6,041	29,110	5,822	5,751
2003	6,577	30,067	6,013	6,339
2004	5,855	32,983	6,597	6,000
2005	6,939	35,007	7,001	6,657
2006	7,571	37,744	7,549	7,297
2007	8,065			7,835
2008	9,314			8,870

Применяя метод полсредних, разбиваем интервал 15 лет на два подынтервала с 1994 по 2001 г. и с 2002 по 2008 г. Суммируя данные за первый период и находя среднюю за восемь лет, получаем первую точку (5; 5,36) — середину между 1997, 1998 гг.

Суммируя данные за 2005—2008 гг., получаем вторую точку (7; 1,94) — 2005 г. Через эти точки проводим прямую тренда.

Как видим, метод определения тренда «на глаз» и способом полусредних дают близкие прямые.

Однако изложенные методы определения тренда хотя и предельно просты, но весьма не точны. Поэтому на практике и особенно в современных условиях широкого распространения компьютерных технологий применяют более совершенные методы.

Простейшим из таких методов является метод скользящих средних. Суть его состоит в том, что выбирается интервал времени определенной длины, включающий чаще всего нечетное число единиц времени, для которых определены уровни ряда. Для интервала вычисляется среднее значение уровня, которое служит новым значением ряда на этом интервале.

Вычисления начинаются с первой точки ряда: интервал как бы прикладывается к оси времени, включая эту точку. После вычисления среднего интервал сдвигается на одну единицу времени и снова вычисляется среднее.

Так продолжается до тех пор, пока не исчерпаются все единицы оси времени. В результате вместо действительных уровней получают последовательность средних величин, которые в итоге сглаживают эмпирические данные ряда.

Поэтому описанный метод часто называют простейшим методом сглаживания. Он дает представление об общей тенденции ряда особенно в рядах с заметными сезонными колебаниями и неочевидным характером тренда.

В табл. 5.2 в колонке 4 показаны результаты применения метода скользящих средних к данным рассматриваемого ряда регистрации автомобилей. В качестве длины интервала выбран отрезок в пять лет, в современной терминологии часто называемый окном. В результате сглаживания получены 11 точек ряда, соединенных пунктирной линией (рис. 5.2).

Определенные длины интервала скользящей средней (размера окна) полностью находятся в руках статистика. Если в ряде явно выражена сезонная компонента, то размер окна необходимо привязывать к периоду сезонности. В других случаях его размер чаще всего выбирают равным трем, пяти или семи единицам времени.

В том случае, когда скользящая средняя все же определяется по четному числу членов ряда, найденное среднее значение определяемого уровня соответствует точке оси времени, находящейся посредине между датами t_i, t_{i+1}. Поэтому средний уровень находится как среднее из двух смежных скользящих средних и относится к соответствующей дате ряда.

Когда тренд ряда имеет явно нелинейный характер и желательно сохранить мелкие колебания значений уровня, применяется более совершенный метод так называемого экспоненциального сглаживания. Суть его состоит в том, что последовательность значений уровней ряда находится по выражению

$$\hat{y}_t = (1 - \alpha)\hat{y}_{t-1} + \alpha y_t,$$

где $\hat{y}_{t-1}$ — полученное в момент времени $t - 1$ по этому выражению теоретическое значение уровня ряда;

y_t — текущее эмпирическое значение уровня ряда;

α — коэффициент экспоненциального сглаживания, $0 < \alpha < 1$.

Степень сглаживания зависит от коэффициента α. Чем он больше, тем сильнее влияние эмпирического значения y_t, и наоборот, чем он меньше, тем сильнее влияние предшествующего теоретического значения уровня y_{t-1}.

В табл. 5.2 результаты экспоненциального сглаживания при $\alpha = 0{,}7$ приведены в последней колонке. При этом первое значение сглаженного уровня y_1 равно исходному значению 5,061.

Второе значение:

$$\hat{y}_2 = 0{,}3 \cdot \hat{y}_1 + 0{,}7 \cdot \hat{y}_2 = 0{,}3 \cdot 5{,}061 + 0{,}7 \cdot 4{,}158 =$$

$$= 1{,}5187 + 2{,}9106 = 4{,}429.$$

Третье значение:

$$\hat{y}_3 = 0{,}3 \cdot \hat{y}_2 + 0{,}7 \cdot y_3 = 0{,}3 \cdot 4{,}429 + 0{,}7 \cdot 5{,}739 =$$

$$= 1{,}329 + 4{,}017 = 5{,}346.$$

Изложенные методы скользящей средней и экспоненциального сглаживания дают лишь представление о тенденции ряда, однако не выражают ее как аналитическую функцию $\hat{y} = f(t)$, при помощи которой, экстраполируя эту функцию, можно дать правдоподобный прогноз поведения ряда в будущем.

Такую возможность предоставляют методы, среди которых наиболее точным в отношении ошибки приближения эмпирических данных теоретической кривой является *метод наименьших квадратов*.

Согласно этому методу сумма квадратов отклонений эмпирических значений ряда от линии регрессии должна быть минимальной. Поэтому вначале необходимо определить линию регрессии, которая бы выражалась в виде аналитической функции.

Опыт показывает, что довольно часто тренды рядов могут изменяться линейно, т. е. усредненные значения уровней ряда расти или убывать с постоянной скоростью, говорят с постоянным темпом. В этом случае аналитической функцией, отображающей такое поведение уровней ряда, является линейная функция

$$\hat{y} = at + b,$$

где α, $b = \text{const}$.

Функцией, отображающей равно ускоренное (равно замедленное) изменение тренда, является функция

$$\hat{y} = a_2 t^2 + a_1 t + b,$$

где a_2 — коэффициент, характеризующий постоянное изменение темпа.

Это так называемая квадратичная функция, согласно которой тренд изменяется как квадрат времени.

Кубическая функция

$$\hat{y} = a_3 t^3 + a_2 t^2 + a_1 t + b$$

характеризует переменное изменение темпа ряда. При этом коэффициент a_3 определяет его ускорение или замедление.

Нередко возможно изменение темпа и по степенной зависимости $\hat{y} = at^x$, и по экспоненциальной $y = a^t$.

В тех случаях, когда отмечается замедление изменения тренда в конце периода, возможно его приближение логарифмической функцией $\hat{y} = a \ln t$.

Когда же налицо плавное падение тренда, его целесообразно отображать обратной функцией $\hat{y} = \dfrac{1}{ax + b}$.

Для того чтобы выбрать наиболее подходящую функцию аппроксимации тренда, необходимо тщательно анализировать исходные графики уровней ряда, а также графики сглаженных значений, полученные методами скользящей средней и экспоненциального сглаживания. Вместе с тем современные компьютерные технологии существенно облегчают задачу, позволяя методом пе-

ребора подобрать наиболее подходящую аналитическую функцию тренда.

Тогда как анализ тенденции осуществляется для определения прогноза на длительный срок времени, анализ сезонной компоненты временного ряда применяется для прогноза на более короткий период.

Например, при составлении прогнозов по сбыту одежды и обуви должны учитываться сезонные колебания в структуре закупок потребителей.

Анализ сезонной компоненты временного ряда отличается от анализа его тенденции двумя моментами.

Во-первых, тенденция определяется непосредственно по исходным данным ряда, а сезонная компонента находится путем исключения из данных части компонент.

Во-вторых, тенденция представляется линией или ее уравнением, а сезонная величина для каждого сезона (месяца, квартала) вычисляется в виде индекса, т. е. числового показателя, выраженного в процентах.

Практически сезонная компонента вычисляется в виде индекса для каждой исследуемой части года. Например, если исследуются ежемесячные колебания, то индексом служит отношение уровня ряда, приходящегося на данный месяц, к средней ряда за весь год, умноженное на 100.

Таким образом, если, например, индекс деловой активности для некоторого месяца равен 100, а для другого 110, это означает, что в первом случае активность равна среднегодовой, а во втором — на 10 % выше.

В качестве средней ряда чаще всего используется скользящая средняя. Поэтому, чтобы определить месячные индексы сезонности, необходимо определить годовую скользящую среднюю.

Предположим, что мы располагаем ежегодной регистрацией автомобилей с 2004 по 2008 г., т. е. за последние пять лет. Требуется определить помесячное влияние сезонных факторов на эту регистрацию.

Согласно методике сначала необходимо определить годовую скользящую среднюю, т. е. найти скользящее среднее количество регистраций автомобилей с января 2004 по декабрь 2004 г., с февраля 2004 по январь 2005 г. и т. д.

После этого каждую скользящую среднюю необходимо разделить на 12, получив тем самым среднее суммы за месяц. Далее

следует отнести число регистраций за каждый месяц к среднемесячным скользящим средним и умножить на 100, определив таким образом индекс сезонности каждого месяца.

На практике эти индексы обычно уточняют, применяя различные методы усреднения, находя среднее за ряд лет, например как среднее арифметическое. Затем индексы корректируют.

В табл. 5.3 приведены результаты расчетов индекса сезонности, полученные описанным методом.

Таблица 5.3. **Индексы сезонности регистрации автомобилей**

Месяц	Год				Уточненный индекс сезонности, %
	2004	2005	2006	2007	
Январь	—	90,14	93,12	92,55	92,2
Февраль	—	80,47	83,39	86,42	83,1
Март	—	100,54	94,69	108,54	100,2
Апрель	—	121,72	119,93	120,95	120,5
Май	—	114,36	116,31	111,53	113,9
Июнь	—	110,16	112,50	109,35	109,7
Июль	105,82	111,46	107,36	—	107,0
Август	92,58	86,67	95,45	—	92,4
Сентябрь	63,86	63,04	81,88	—	63,6
Октябрь	114,40	111,09	93,97	—	110,7
Ноябрь	106,12	98,74	79,86	—	98,4
Декабрь	106,10	102,36	106,26	—	105,7

Из табл. 5.3 (последняя колонка) следует, что сезонный пик числа регистрации новых автомобилей приходится на апрель, а сезонный минимум — на сентябрь. Данные результаты относительно правдоподобны, так как весной обычно покупают новые автомобили, а после отпуска их стараются продать. Пик регистрации в октябре можно объяснить приобретением новых моделей автомобилей.

В то время как анализ тренда имеет прямую практическую ценность для определения длительного прогноза, а анализ сезонной компоненты для определения прогноза на короткий период времени, анализ циклической и нерегулярной компоненты для менеджеров имеет существенно меньшую ценность.

Тем не менее статистика располагает методами такого анализа. Сущность классического подхода к выявлению влияния циклической и нерегулярной компонент состоит в исключении влияния тенденции и сезонной компоненты на уровни временного ряда. В связи с этим классический метод часто называют остаточным.

5.2. Индексы

Наряду с анализом временных рядов в статистике применяются построение и анализ индексов, т. е. числовых показателей, зависящих от времени.

Широко известны индексы потребительских и оптовых цен, промышленного производства и др.

Индекс всегда показывает результат сравнения некоторой абсолютной величины в данный момент с той же величиной, измеренной в некоторый предшествующий момент, который был принят как базовый. Например, если в качестве базового года регистрации количества автомобилей принять 1994 г., то отношения регистраций количеств автомобилей во все последующие годы, включая 2008 г., к базовому количеству 1994 г., умноженные на 100, будут составлять индексы продаж автомобилей.

Различают индексы простые и сложные. *Простой индекс* выражает изменение по отношению к базе одного товара. Например, индекс цен на автомобиль марки «Форд» с объемом цилиндра 2 л представляет собой простой индекс. В то же время индекс потребительских цен на некоторую группу товаров (сахар, масло, хлеб, мясо) является *сложным индексом*.

Другая классификация индексов — это индексы количества, цен и стоимости. Например, индекс промышленного производства представляет собой *индекс количества*. Индекс потребительских или оптовых цен — это *индекс цен*. Индекс запасов, где не различают цену и количество, представляет *индекс стоимости*.

Расчет простых индексов не сложен. Если C_0 — цена товара в базовом периоде, а C_1 — в последующем заданном периоде, то простой индекс цен, т. е. относительная цена, определяется по выражению

$$I_c = \frac{C_1}{C_0} \cdot 100. \qquad (5.1)$$

Например, если среднюю цену свежего мяса (говядины), равную 250 руб. за 1 кг в 2005 г., принимают в качестве базы, а средняя его цена в 2008 г. составляла 320 руб., то индекс цен мяса по выражению (5.1) равен $I_c = \frac{320}{250} \cdot 100 = 128\ \%$.

Аналогично, если Q_0 — количество товара в базовом периоде, а Q_1 — в рассматриваемом, то индекс цен количества определяется по выражению

$$I_Q = \frac{Q_1}{Q_0} \cdot 100. \qquad (5.2)$$

Например, если в 2005. базовом году один человек европейской части России употреблял 36 кг мяса в год, а в 2008 г. объем этого мяса составил 30 кг, то индекс количества потребляемого мяса составил $I_Q = \frac{30}{36} \cdot 100 = 83{,}3\ \%$.

С учетом цен на мясо и объема его потребления в базовом 2005 г. индекс стоимости определяется по выражению

$$I_{QC} = \frac{C_1 Q_1}{C_0 Q_0} \cdot 100. \qquad (5.3)$$

Согласно этому $I_{QC} = \frac{320 \cdot 30}{250 \cdot 36} \cdot 100 = 180\ \%$.

В коммерческой и экономической областях деятельности большинство рассчитываемых индексов являются сложными. Расчет сложного индекса, т. е. индекса, характеризующего группу товаров, требует предварительного отбора товаров и включения его в группу с определенным весом.

Например, если житель некоторого региона потребляет 15 кг мяса в месяц и покупает в год три пары туфель, т. е. $\frac{1}{4}$ пары в ме-

сяц, то в месячный индекс потребительских цен на мясо и обувь цена мяса должна включаться с весом 1,5, а цена туфель — с весом 0,25.

Основной базой для определения весов является количество каждого товара, проданного за определенный период, и эти количества используются для определения сложных индексов цен как базисного, так и текущего года.

Пусть $i = 1, 2, \ldots, n$ — типы товаров, включенных в группу;

$Q_{01}, Q_{02}, \ldots, Q_{0i}, \ldots, Q_{0n}$ — количества товаров, проданных в базисный год;

$C_{01}, C_{02}, \ldots, C_{0i}, \ldots, C_{0n}$ — цены на товары базисного года;

$C_{11}, C_{12}, \ldots, C_{1i}, \ldots, C_{1n}$ — цены на товары текущего года;

$Q_{11}, Q_{12}, \ldots, Q_{1i}, \ldots, Q_{1n}$ — количества товаров, проданных в текущий год.

Тогда, если в качестве весов взять количество базисного года, сложный индекс цен будет определяться по формуле

$$I_L = \frac{\sum_{i=1}^{n} C_{1i} Q_{0i}}{\sum_{i=1}^{n} C_{0i} Q_{0i}}. \qquad (5.4)$$

Этот индекс называется *индексом Ласпейреса*.

Вместо количеств продаж базисного года Q_{0i}, $i = 1, 2, \ldots, n$ можно взять количества продаж текущего года Q_{1i}, $i = 1, 2, \ldots, n$. Тогда получим сложный индекс

$$I_P = \frac{\sum_{i=1}^{n} C_{1i} Q_{1i}}{\sum_{i=1}^{n} C_{0i} Q_{1i}}, \qquad (5.5)$$

называемый *индексом Пааше*.

Таким образом, при вычислении индекса Ласпейреса осуществляется взвешивание базовых и текущих цен на основании количеств товаров, проданных в базовый период. Наоборот, вычисление индекса Пааше предусматривает взвешивание цен

на товары на основании их количеств, проданных в текущий период.

Индекс Пааше сравнивает цену комплекта товаров данного периода с ценой, которую имел бы этот же комплект в базисном периоде.

Индекс Ласпейреса сравнивает цену товара в новом периоде с ценой в базисном периоде, используя базисную цену.

В связи с изменением содержания потребительской корзины с течением времени использование ее базисной цены, т. е. расчеты индекса Ласпейреса, приводит к ошибке в определении стоимости жизни.

С другой стороны, снижение цены на некоторый товар из этой корзины приводит к его интенсивному потреблению, которое фиксируется при расчете индекса Пааше. Поэтому такой индекс не отражает реальную ситуацию.

Какой индекс лучше как показатель уровней изменения цен и стоимости, представляет собой предмет не статистического анализа.

5.3. Компьютерные технологии анализа временных рядов

Пакет **Анализ данных** содержит две процедуры сглаживания временных рядов и таким образом определения характера тренда. Это процедуры **Скользящее среднее** и **Экспоненциальное сглаживание**.

Вызов процедуры **Скользящее среднее** осуществляется из окна **Инструменты анализа** выделением этой процедуры и нажатием кнопки **ОК**. Для выполнения процедуры необходимо в открывшемся окне заполнить поле *входной интервал*, для чего ввести ссылки на ячейки листа Excel, в которых хранятся уровни ряда.

Далее в поле *интервал* необходимо ввести размер интервала сглаживания, отметить выходной интервал и вывод графика, а также стандартных погрешностей. После этого следует нажать кнопку **ОК**.

В результате выполнения процедуры на экран будет выведена таблица с исходными и сглаженными уровнями, а также стандартными погрешностями. Кроме этого, будет выведен гра-

фик фактических и сглаженных значений. По этому графику можно проследить характер тренда.

Процедура **Экспоненциальное сглаживание** вызывается аналогично процедуре **Скользящее среднее**. В открывшемся окне процедуры, кроме поля *входной интервал*, необходимо заполнить поле *фактор затухания*, в которое вводится значение α, например 0,3. Инициируются также поля *вывод графика* и *стандартные погрешности*. После этого нажимается кнопка **ОК**.

В результате выполнения процедуры на экран выводится таблица исходных и сглаженных уровней ряда, стандартных погрешностей и график, по которому можно судить о характере тренда.

Построение эмпирических формул тренда осуществляется в режиме **Мастер диаграмм**. Для этого в ячейки листа Excel предварительно вводятся исходные данные ряда, например уровни продажи по годам, представленные в табл. 5.4 [5].

Таблица 5.4. **Объемы продаж**

Годы	2005	2006	2007	2008	2009
Объем продаж	16,4	17,04	17,24	18,57	19,08

Далее из пункта **Вставка** главного меню Excel щелчком мыши вызывается **Мастер диаграмм**, выбирается точечная диаграмма и путем последовательного нажатия кнопки **Готово** строится первоначальный график тренда.

По графику судят о том, какая функция наиболее подходит для аппроксимации тренда. Практическую помощь при этом получают, используя режим **Добавить линию тренда**.

Для выхода в этот режим необходимо в главном меню сначала щелкнуть пункт **Диаграмма**, а затем в раскрывшемся меню выделить и щелкнуть пункт **Добавить линию тренда**. В результате на экране появится окно с графиками функций: линейная, логарифмическая, полиномиальная, степенная и экспоненциальная.

Необходимо выбрать наиболее подходящую функцию и щелкнуть вкладку окна **Параметры**.

Далее в очередном окне щелчком мыши необходимо активизировать поля: *показать уравнение на диаграмме*, *поместить на диаграмму величину достоверности аппроксимации* $R_d \wedge 2$. После этого следует нажать клавишу **Готово**.

В результате на экран будут выведены исходный и аппроксимирующий графики тренда, уравнение $\hat{y} = f(x)$ и коэффициент детерминации R_d^2.

Судят о качестве приближения по значению коэффициента детерминации R_d^2. Чем он ближе к единице, тем более точная аппроксимация, т. е. тем более точно полученное уравнение приближает тренд. Поэтому на практике поиск самого подходящего уравнения осуществляют методом его последовательного подбора по коэффициенту R_d^2. Лучшей аппроксимирующей кривой будет та, для которой R_d^2 максимально.

Таким образом, процедура подбора сводится к последовательному построению при помощи **Мастера диаграмм** исходной кривой, выбора функции аппроксимации и построения кривой тренда с показом уравнения $y = f(x)$ и коэффициента детерминации R_d^2.

Контрольные вопросы

1. Что представляет собой временной ряд? Приведите примеры таких рядов.
2. Зачем производится анализ временных рядов?
3. Какие основные параметры характеризуют временные ряды? Как они называются?
4. Как классифицируются временные ряды?
5. Какие факторы выделяют при анализе временных рядов?
6. Что представляют собой тенденция, сезонная, циклическая и неопределенная компоненты ряда?
7. К чему сводится классический метод анализа временных рядов?
8. Перечислите методы выявления тренда временных рядов.
9. В чем смысл метода полусредних, скользящих средних и экспоненциального сглаживания?
10. Какой метод применяется для построения аналитической функции, приближающей тренд?

11. Какие функции чаще всего применяются для приближенной замены эмпирических данных тренда? Перечислите функции и дайте им объяснение.
12. Какие компьютерные средства пакета **Анализ данных** используются для сглаживания рядов? Какие результаты позволяют получить применение этих средств?
13. Каков порядок подбора на компьютере наиболее подходящей аналитической функции, аппроксимирующей тренд? По какому признаку определяется наиболее подходящая функция?
14. С какой целью проводится анализ сезонной компоненты ряда?
15. В чем принципиальное отличие анализа сезонной компоненты от анализа тренда?
16. В каких величинах принято представлять сезонную компоненту?
17. Можно ли считать сезонной компонентой некоторого месяца отношение уровня ряда за данный месяц к средней ряда за весь год?
18. Какая средняя чаще всего используется при определении месячных индексов сезонности?
19. Какие индексы экономической и коммерческой деятельности вам известны?
20. Что показывают индексы оптовых и потребительских цен, индекс промышленного производства?
21. Как классифицируются индексы?
22. Как рассчитываются простые индексы цен и количество?
23. Как рассчитываются простые индексы стоимости?
24. Как рассчитывается индекс Ласпейреса?
25. Как рассчитывается индекс Пааше?
26. В чем принципиальное различие индексов Ласпейреса и Пааше?

Задачи

1. Годовые объемы продаж продукции магазинами торговой сети представлены в табл. 5.5.

Таблица 5.5. **Годовые объемы продаж продукции, млн руб.**

Годы	2003	2004	2005	2006	2007	2008
Объемы продаж	1,5	1,3	1,1	1,7	1,9	2,3

Постройте график продаж, проведите линию тенденции «на глаз». Проведите линию тенденции при помощи метода полусредних, сглаживание данных ряда методом скользящих средних. Для этого используйте соответствующую процедуру пакета **Анализ данных**.

Проведите сглаживание данных ряда методом экспоненциального сглаживания.

Подберите аналитическую функцию, наиболее точно приближающую тенденцию ряда.

На основе функции дайте прогноз объема продаж на 2009 и 2010 гг.

2. Квартальные объемы продаж представлены в табл. 5.6.

Таблица 5.6. **Квартальные объемы продаж, тыс. руб.**

Год	Квартал	Объем продаж
2005	I	175
	II	263
	III	326
	IV	297
2006	I	247
	II	298
	III	366
	IV	341

Окончание табл. 5.6

Год	Квартал	Объем продаж
2007	I	420
	II	441
	III	453
	IV	399
2008	I	426
	II	449
	III	482
	IV	460

Методом скользящих средних с интервалом 4 найдите сглаженные уровни продаж по кварталам 2005—2008 гг.

Выведите график сглаживания и определите тенденцию.

Подберите наиболее подходящую аналитическую функцию, приближающую тренд, и выведите на экран ее график.

3. В табл. 5.7 приведены данные по производству сливочного масла и твердого сыра. Таблица 5.8 содержит оптовые цены продукции в 2005—2008 гг.

Таблица 5.7. **Производство масла и сыра**

Продукт	2005 г.	2006 г.	2007 г.	2008 г.
Масло	614,88	567,84	576,96	534,84
Сыр	634,08	652,32	690,6	697,28

Таблица 5.8. **Цены масла и сыра**

Продукт	2005 г.	2006 г.	2007 г.	2008 г.
Масло	110	115	118	125
Сыр	230	235	244	252

Взяв в качестве базового 2005 г., вычислите индексы цен масла и сыра для указанных четырех лет.

Что можно сказать о процентных изменениях в оптовых ценах на масло и сыр в течение 2005—2008 гг.?

Вычислите индексы количества для масла и сыра, взяв в качестве базисного 2005 г.

Что можно сказать о процентных изменениях количества производства масла и сыра за 2005—2008 гг.?

Вычислите стоимость производства масла и сыра в 2005—2008 гг.

Вычислите индексы Ласпейреса и Пааше для 2008 г., взяв в качестве базисного 2005 г.

СТАТИСТИЧЕСКИЕ РЕШЕНИЯ

Глава 6
ЭЛЕМЕНТЫ ТЕОРИИ СТАТИСТИЧЕСКИХ РЕШЕНИЙ

6.1. Классификация задач принятия решений

Изложенный в пяти главах материал представляет собой классические методы статистического анализа, для проведения которого требовались те или иные наборы числовых данных, получаемые либо путем наблюдений, либо проведением экспериментов.

Статистические заключения, в основе которых лежат классические методы, предполагают использование элементов теории вероятностей, когда вероятность рассматривается как объективная числовая мера исхода случайного события. В результате менеджер, опираясь на статистический вывод, может принять решение с известной долей риска, определяемой величиной вероятности.

Однако во многих ситуациях управляющему приходится принимать единичные решения, используя при этом некоторые предположения относительно того, как будут развиваться события, или опираясь на аналогии. В этих случаях для использования хоть какой-либо числовой меры для принятия решения применяют понятие не объективной, а субъективной вероятности. Такая вероятность интерпретируется как степень уверенности лица, принимающего решение, о том, что рассматриваемое со-

бытие произойдет. Кроме того, используют меру оценки полезности возможного варианта решения.

По современным воззрениям принятие решения в сложных ситуациях представляет собой проблему, включающую следующие составляющие.

1. Должна быть четко и корректно сформулирована цель, т. е. однозначно оговорен результат, для достижения которого принимается решение.

2. Лицо, принимающее решение, несет полную ответственность за результат решения.

3. В обязательном порядке должны иметь место различные варианты достижения цели. Иными словами, должен быть обеспечен выбор различных путей ее достижения. Когда этих атрибутов нет, выбор один: принять решение или отклонить.

4. В соответствии с вариантами выбора путей достижения цели должны быть определены исходы их выбора, измеренные в тех или иных единицах.

5. Должны быть определены все факторы, влияющие на исходы решений, т. е. установлено воздействие внешней среды.

6. Необходимо наличие правил, в соответствии с которыми выбирается тот или иной вариант решения.

В зависимости от того, насколько информирован принимающий решение и как определены исходы выбираемых вариантов, принята следующая классификация задач принятия решений: 1) в условиях определенности; 2) в условиях риска; 3) в условиях неопределенности; 4) в условиях конфликта или противодействия противника.

Условно схема принятия решений представлена в табл. 6.1.

Таблица 6.1. **Графическое представление задачи принятия решений**

Варианты решений	Исходы решений			
	1	2	...	n
V_1	l_{11}	l_{12}	...	l_{1n}
V_2	l_{21}	l_{22}	...	l_{2n}
....	...	...	...	...
V_m	l_{m1}	l_{m2}	...	l_{mn}

По существу, мы имеем прямоугольную таблицу — матрицу, колонки которой обозначены как исходы I, а строки — как варианты решений V.

В клетках таблицы, образуемых пересечением строк и столбцов, указаны результаты исходов I_{ij}, соответствующих выбору тех или иных вариантов. Эти результаты могут представлять собой либо потери, либо доходы и в общем случае называются полезностями.

Таким образом, полезность, например, I_{11} обозначает некоторую величину, которая достигается при выборе варианта V_1 и исходе I_1.

Принятие решений в условиях определенности характеризуется однозначной связью между принятым вариантом решения и его исходом. Иными словами, управляющий точно знает, что, выбрав такой-то вариант решения, он получит заведомо известный исход.

Применительно к приведенной схеме решений это означает, что каждому варианту V_i, $i = 1, 2, \ldots, m$ будет соответствовать один и только один элемент матрицы, I_{ij}.

Поэтому принятие решений в условиях определенности сводится к тому, чтобы выбрать вариант V_i, доставляющий наибольшую полезность I_{ij}.

В практических задачах такого содержания количество вариантов чрезвычайно велико. Поэтому, привлекая математику, задачу о выборе лучшего решения в общем случае формулируют как задачу о поиске наибольшего значения функции $L = f(v)$, представляющей зависимость полезности L от варианта V.

Такие задачи получили название задач математического программирования (линейного, нелинейного, дискретного), и их освещение, а также анализ не является предметом статистики.

6.2. Принятие решений в условиях риска

Эта задача возникает в том случае, когда каждый вариант решений V_i, $i = 1, 2, \ldots, m$ имеет несколько исходов I_j, $j = 1, 2, \ldots, n$, объективная или субъективная вероятность каждого из которых может быть определена.

Рассмотрим пример. Некто, выходя из дома на длительную прогулку в лес, должен надеть одежду, соответствующую погоде.

Возможны три состояния погоды: солнечно, переменно и дождь. В соответствии с этим следует надеть: 1) спортивный костюм; 2) спортивный костюм, шапочку и ветровку; 3) спортивный костюм, шапочку, резиновые сапоги и взять зонт.

Таким образом, в данной ситуации имеем три варианта решения относительно выбора одежды, каждому из которых соответствует три состояния погоды (три исхода). Поэтому задача выбора решения представляется квадратной матрицей (табл. 6.2).

Таблица 6.2. **Матрица вариантов решений, исходов и полезностей**

Одежда	Погода (исходы)		
	солнечно	переменно	дождь
Спортивный костюм	1	0,4	2
Спортивный костюм, ветровка	0,4	1	−1,5
Спортивный костюм, зонт	0,3	0,5	1

В клетках матрицы, находящихся на пересечении ее строк и столбцов, проставлены полезности, которые могут быть получены, если вариант решения соответствует или не соответствует состоянию погоды. Эти полезности выбраны произвольно из тех соображений, что в том случае, когда вариант выбора одежды соответствует погоде, полезность максимальна, а когда лицо, гуляющее в спортивном костюме, попадает под дождь, она минимальна.

Однако, прежде чем выйти на прогулку, спортсмен смотрит на барометр, который может показывать солнечно, переменно или дождь. Из прежних показателей барометра известно, что доверять его показаниям можно примерно на 70 %.

Предположим, в данном случае барометр показывает солнечно. Тогда вероятность P_s того, что погода будет именно такой, равна 0,7. Вероятности того, что может быть переменно или дождь, имея в виду показания барометра, определяются как $P_p = 0{,}15$. $P_d = 0{,}15$. В результате получим матрицу вероятностей исходов, представленную в табл. 6.3.

Располагая матрицами полезностей и вероятностей исходов, можно с определенной долей риска найти наиболее подходящий вариант выбора одежды.

Таблица 6.3. **Матрица вероятностей исходов**

Вероятности		
P_s	P_p	P_d
0,7	0,15	0,15
0,7	0,15	0,15
0,7	0,15	0,15

Для принятия решений в условиях риска обычно используют так называемую ожидаемую полезность каждого варианта решения. Она вычисляется как сумма попарных произведений полезностей на вероятности для каждой строки матрицы принятия решений (см. табл. 6.2). В данном случае, выполняя указанные операции, получим следующие ожидаемые полезности.

Спортивный костюм:

$$1 \cdot 0{,}7 + 0{,}4 \cdot 0{,}15 - 2 \cdot 0{,}15 = 0{,}7 + 0{,}06 - 0{,}3 = 0{,}46.$$

Спортивный костюм, ветровка:

$$0{,}4 \cdot 0{,}7 + 1 \cdot 0{,}15 - 1 \cdot 0{,}15 = 0{,}28 + 0{,}15 - 0{,}15 = 0{,}28.$$

Спортивный костюм, зонт:

$$0{,}3 \cdot 0{,}7 + 0{,}5 \cdot 0{,}15 + 0{,}15 = 0{,}21 + 0{,}075 + 0{,}15 = 0{,}435.$$

Когда ожидаемые полезности вычислены, можно выбрать лучший вариант решения проблемы. Для этого используют следующее правило: выбрать тот вариант, ту стратегию, для которого ожидаемая полезность наибольшая (максимальная).

В данном случае такой стратегией является «надеть спортивный костюм», так как среди всех ожидаемых полезностей (0,46; 0,28; 0,435) значение 0,46 максимально и соответствует первому варианту принятия решения.

Трудности принятия решений в условиях риска состоят в том, чтобы более-менее правдоподобно определить полезности и вероятности исходов. Эта задача полностью ложится на лицо, принимающее решение. При этом, если задание полезностей чисто субъективно, для определения вероятностей исходов могут

быть использованы вспомогательные сведения: в нашем случае показания барометра и степень доверия к этим показаниям.

В некоторых задачах для определения вероятностей допустимо проведение эксперимента. В следующем примере будет показано, как более объективно определяются полезности.

Каждый летний сезон начальник крупной железнодорожной станции в связи с ожидаемым наплывом отдыхающих принимает решение о прицеплении к поездам, идущим к морю, некоторого количества дополнительных вагонов. При этом он располагает сведениями, что каждый дополнительный вагон в месяц приносит средний доход 1,5 млн руб., а эксплуатация вагона составляет 600 тыс. руб.

Таким образом, прибыль (полезность) от использования одного вагона, заполненного сезонными пассажирами, составляет 1 500 000 – 600 000 = 900 000 руб.

В том случае, когда в эксплуатацию пущено, например, 5 вагонов, а сезонными пассажирами заполнено только два вагона, получаем полезность 1 500 000 · 2 – 600 000 · 5 = 0.

Используя эту методику, экономический отдел управляющего составил матрицу полезностей, представленную в табл. 6.4.

Таблица 6.4. **Матрица полезностей эксплуатации вагонов, тыс. руб.**

Число прицепляемых вагонов (варианты решений)	Число требуемых вагонов (исходы)					
	0	1	2	3	4	5
0	0	0	0	0	0	0
1	–600	900	900	900	900	900
2	–1200	300	1800	1800	1800	1800
3	–1800	–300	1200	2700	2700	2700
4	–2400	–900	600	2100	3600	3600
5	–3000	1500	0	500	3000	4500

Если число прицепных вагонов, например, 1, а требуемых 0, то согласно методике расчета получаем полезность, равную 600 тыс. руб., т. е. убыток. Это число представлено на пересече-

нии второй строки и первого столбца матрицы. Если число прицепленных вагонов 5, а требующихся 5, получаем доход 1,5 млн руб. · 5 – 0,6 млн руб. · 5 = 4,5 млн руб.

Это число представлено на пересечении пятой строки матрицы и пятого ее столбца.

Из предшествующих сезонов работы станции известны вероятности требуемых вагонов для перевозки пассажиров. Они представлены в табл. 6.5.

Таблица 6.5. **Вероятности потребности вагонов**

Число требуемых вагонов	0	1	2	3	4	5
Вероятность	0	0,1	0,2	0,3	0,3	0,1

Теперь для того, чтобы найти лучший вариант решения (сколько прицеплять вагонов), необходимо для каждого варианта вычислить ожидаемую полезность.

Имеем:

Вариант 1:

$$900 \cdot 0{,}1 + 900 \cdot 0{,}2 + 900 \cdot 0{,}3 + 900 \cdot 0{,}3 + 900 \cdot 0{,}1 = 900.$$

Вариант 2:

$$300 \cdot 0{,}1 + 1800 \cdot 0{,}2 + 1800 \cdot 0{,}3 + 1800 \cdot 0{,}3 + 1800 \cdot 0{,}1 = 1650.$$

Вариант 3:

$$-300 \cdot 0{,}1 + 1200 \cdot 0{,}2 + 2700 \cdot 0{,}3 + 2700 \cdot 0{,}3 + 2700 \cdot 0{,}1 = 2100.$$

Вариант 4:

$$-900 \cdot 0{,}1 + 600 \cdot 0{,}2 + 2100 \cdot 0{,}3 + 3600 \cdot 0{,}3 + 3600 \cdot 0{,}1 = 2100.$$

Вариант 5:

$$1500 \cdot 0{,}1 + 0{,}02 + 500 \cdot 0{,}3 + 3000 \cdot 0{,}3 + 4500 \cdot 0{,}1 = 1350.$$

Таким образом, максимальная полезность достигается одновременно для третьего и четвертого вариантов. Это означает, что можно использовать три или четыре прицепных вагона для перевозки сезонных пассажиров.

6.3. Принятие решений в условиях неопределенности

Задачи такого рода характеризуются тем, что принимающий решение менеджер может определить исходы для варианта решения, однако ни объективные, ни субъективные вероятности этих исходов не поддаются вычислению. Таким образом, для принятия решения управляющий располагает только матрицей полезностей (табл. 6.4).

Теория решений для выбора наиболее подходящего варианта решения в таких условиях предлагает несколько правил, которые носят имена разработавших их авторов.

Правило Вальда, часто называемое критерием Вальда, состоит в том, что для каждого варианта решения V_i, $i = 1, 2, \ldots, m$ находится наименьшая полезность l_{ij}, а затем среди найденных полезностей выбирается та, которая максимальна. Она и определяет вариант решения.

Таким образом, практически сначала в каждой строке матрицы находится минимальный элемент. Затем среди этих элементов находится максимальный элемент, который указывает строку матрицы, в которой он стоит.

Этот критерий носит название максимина и обеспечивает некоторую гарантированную полезность при самом худшем исходе. Он соответствует логике пессимиста, который предполагает, что обязательно произойдет самое худшее.

Правило Гурвица, или критерий Гурвица, состоит в том, что делается предположение о возможном худшем и лучшем исходах. Полагают, что лучший исход может произойти с вероятностью α, а худший с $1 - \alpha$.

Правило состоит в том, что сначала в каждой строке матрицы полезностей находят максимальную полезность $\max l_{ij}$ и минимальную полезность $\min l_{ij}$. Затем находят взвешенную по вероятностям их сумму $\alpha \cdot \max l_{ij} + (1 - \alpha) \cdot \min l_{ij}$. После чего находят элемент матрицы, который равен максимальному значению взвешенной суммы. Тем самым определяют строку матрицы, соответствующую этому элементу, и, следовательно, вариант решения.

В том случае, когда $\alpha = 0$, правило Гурвица равнозначно правилу Вальда. Когда же $\alpha = 1$, получаем стратегию, которая определена максимальной полезностью из максимальных значе-

ний каждой строки. Это стратегия оптимиста, который уверен, что реализуется самый лучший исход.

Правило Лапласа предполагает, что все исходы равновероятны. Тогда выбор варианта решения согласно этому правилу осуществляется аналогично выбору решения в условиях риска. Практически для каждой строки матрицы находится ожидаемая полезность и среди найденных значений выбирается наибольшее. Оно и определяет строку матрицы, т. е. вариант решения. При этом вероятности исходов определяются как $P = \frac{1}{n}$, где n — количество исходов.

Правило Севиджа, или критерий Севиджа, предполагает поиск варианта решения, который минимизирует так называемые максимальные сожаления. При этом под сожалениями подразумевают потери полезностей каждого исхода относительно максимальной полезности.

Сожаления вычисляются следующим образом. В каждом столбце матрицы полезностей находят максимальный элемент (максимальную полезность) и вычитают из него остальные элементы столбца, в том числе и максимальный элемент. В результате получают матрицу сожалений, другими словами, матрицу возможных полезностей.

Правило Севиджа используется применительно к этой матрице. Для этого в каждой строке матрицы находится максимальный элемент и далее среди максимальных элементов определяется наименьший из них, который и указывает строку матрицы, т. е. вариант решения.

Таким образом, критерий Севиджа представляет собой минимизацию максимальных потерь — минимакс потерь.

Продемонстрируем применение рассмотренных правил на следующей задаче.

Фирма разработала и освоила технологию изготовления телевизоров с экраном на светодиодах. Широкое потребление таких телевизоров в течение последующих 10 лет считается достоверным. Однако вероятности, как будут продаваться телевизоры по годам, неизвестны.

Руководство фирмы считает, что возможна следующая политика производства и продаж: немедленное производство и сбыт; ограничение производства; ограничение производства через 2 года; ограничение производства через 5 лет.

Относительно потребления товара опрос потребителей показал, что возможно немедленное интенсивное потребление, интенсивное потребление через 2 года, интенсивное потребление через 5 лет и интенсивное потребление через 8 лет. Требуется выбрать наиболее подходящий вариант производства и сбыта телевизоров.

Экономический отдел фирмы подготовил матрицу полезностей, представленную в табл. 6.6.

Таблица 6.6. **Полезности производства телевизоров**

Варианты решений	Интенсивное потребление (исходы)			
	Немедленное	Через 2 года	Через 5 лет	Через 8 лет
Немедленное производство	80	40	–10	–50
Ограничение производства	30	40	30	10
Ограничение производства через 2 года	20	30	40	15
Ограничение производства через 5 лет	5	10	30	30

Согласно **правилу Вальда**, отражающему взгляд пессимиста, наименьшие полезности для каждого варианта равны –50, 10, 15 и 5, а наибольшая из них равна 15. Она указывает на третью строку матрицы, т. е. на то, что необходимо выбрать вариант, ограничивающий производство и сбыт через 2 года.

Для применения **правила Гурвица** необходимо определить значение α. Выберем его равным 0,7. Тогда $1 - \alpha = 0{,}3$.

Для первой строки матрицы полезностей минимальная полезность равна –50, максимальная 80. Поэтому взвешенная по вероятности полезность первой строки равна $\alpha \cdot 80 + (1 - \alpha)(-50) = 0{,}7 \cdot 80 + 0{,}3(-50) = 56 - 15 = 41$.

Для второй строки матрицы полезностей минимальная полезность равна 10, максимальная 40. Откуда взвешенная полезность второй строки составит $0{,}7 \cdot 40 + 0{,}3 \cdot 10 = 28 + 3 = 31$.

Для третьей строки матрицы полезностей минимальная полезность равна 15, максимальная 40. Поэтому взвешенная полезность равна $0{,}7 \cdot 40 + 0{,}3 \cdot 15 = 28 + 4{,}5 = 32{,}5$.

Для четвертой строки матрицы полезностей минимальная полезность равна 5, максимальная 30. На этом основании взвешенная полезность равна $0,7 \cdot 30 + 0,3 \cdot 5 = 21 + 1,5 = 22,5$.

Максимальная полезность среди взвешенных полезностей равна 41. Она соответствует первой строке матрицы. Таким образом, в соответствии с правилом Гурвица и принятым значением α необходимо выбрать первый вариант производства.

Согласно **правилу Лапласа** вероятности исходов P для этой задачи равны 0,25. На этом основании ожидаемые полезности вариантов решений такие:

$80 \cdot 0,25 + 40 \cdot 0,25 - 10 \cdot 0,25 - 50 \cdot 0,25 = 15;$

$30 \cdot 0,25 + 40 \cdot 0,25 + 30 \cdot 0,25 + 10 \cdot 0,25 = 27,5;$

$20 \cdot 0,25 + 30 \cdot 0,25 + 40 \cdot 0,25 + 15 \cdot 0,25 = 26,75;$

$5 \cdot 0,25 + 10 \cdot 0,25 + 30 \cdot 0,25 + 30 \cdot 0,25 = 18,75.$

Откуда максимальная ожидаемая полезность равна 27,5. Она определяет второй вариант решения, т.е ограничение производства.

Для того чтобы выбрать вариант решения по **правилу Севиджа**, необходимо перейти к матрице сожалений. Она будет иметь вид, представленный в табл. 6.7.

Таблица 6.7. **Матрица сожалений**

Варианты решений	Интенсивное потребление (исходы)			
	немедленное	через 2 года	через 5 лет	через 8 лет
Немедленное производство	0	0	50	80
Ограничение производства	50	0	10	20
Ограничение производства через 2 года	60	10	0	15
Ограничение производства через 5 лет	75	30	10	0

Элементы каждого столбца этой матрицы получены как разности между максимальным элементом, им самим и другими элементами.

Теперь согласно критерию Севиджа в каждой строке матрицы находим максимальный элемент. В результате получаем 80, 50, 60, 75. Из этих элементов выбираем минимальный 50, указывающий на то, что нужно выбрать второй вариант решения — ограничение производства.

Таким образом, три правила (Вальда, Лапласа и Севиджа) указывают на то, что необходимо выбрать второй вариант решения.

В том случае, когда согласно всем правилам были бы выбраны различные варианты решений, окончательный выбор варианта решения чрезвычайно затруднен. Он может быть осуществлен на базе более глубокого анализа рынка, а также квалификации и производственного опыта управляющего фирмой.

Из всех ситуаций, требующих принятия решения, конфликтная ситуация, когда действиям принимающего решение противодействует разумный конкурент, является наиболее сложной. Описание таких ситуаций относится к области теории игр [11].

Контрольные вопросы

1. Какие составляющие определяют проблему принятия решений в сложных ситуациях?
2. Как классифицируются задачи принятия решений?
3. Какими факторами определяется классификация?
4. Как графически представляется общая схема принятия решений?
5. Какие факторы определяют задачи принятия решений в условиях определенности?
6. В чем особенности принятия решений в условиях риска?
7. Какое правило применяется для выбора лучшего варианта решения в условиях риска?
8. Как могут определяться вероятности исходов в задачах принятия решений в условиях риска?
9. Чем отличаются задачи принятия решений в условиях неопределенности от задач принятия решений в условиях риска?

10. Какие правила применяются для выбора лучших решений в условиях неопределенности?
11. Что означают правила максимина и минимакса?
12. Как построить матрицу сожалений?

Задачи

1. Руководство фирмы, производящей продукты питания, нуждается в составлении плана производства и продаж их на рынке. В течение месяца она может произвести и поставить в магазины 0, 1, 2 и 3 т продуктов. Продажа 1 т продуктов приносит доход 315 000 руб. Если продукция не будет продана, убыток составит 420 000 руб.

   Составьте матрицу полезностей производства и продаж 0, 1, 2 и 3 т продукции.

   Вероятности продаж 0, 1, 2 и 3 т продукции соответственно равны 10, 40, 30 и 20 %. Найдите лучший вариант производства продукции.

   Для сформулированной задачи вероятности продаж неизвестны.

   Определите решение по правилам Вальда, Гурвица при $a = 0{,}6$, Лапласа, Севиджа.

   Используя полученные решения, выберите лучший вариант производства продукции.

Литература

1. *Роберт А. Доннели.* Статистика. М., 2006.
2. *Вентцель Е. С.* Теория вероятности (Первые шаги). М., 1977.
3. *Закс Лотар.* Статистическое оценивание. М., 1976.
4. *Львовский Е. Н.* Статистические методы построения эмпирических формул. М., 1988.
5. *Макарова Н. В., Трофимец В. Я.* Статистика в Excel. М., 2002.
6. Теория статистики: учебник / под ред. Р. А. Шмойловой. М., 2001.
7. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М., 1973.
8. *Зайченко Ю. П.* Исследование операций. Киев, 1975.
9. *Чернов Г., Мозес П.* Элементарная теория статистических решений. М., 1962.
10. *Моррис Г.* Оптимальные статистические решения. М., 1974.
11. *Льюс Р., Райфа Х.* Игры и решения. М., 1961.
12. *Кремер Н. Ш.* Теория вероятности и математическая статистика. М., 2009.
13. *Балдин К. В., Башлыков В. Н., Рукосуев А. В.* Теория вероятности и математическая статистика. М., 2010.
14. *Палий И. А.* Прикладная статистика. М., 2009.
15. *Лысенко С. Н., Дмитриева И. А.* Общая теория статистики. М., 2008.

Оглавление

Канцедал Сергей Андреевич

Основы статистики

Учебное издание

Редактор *М. А. Кутепова*
Корректор *О. Н. Картамышева*
Компьютерная верстка *И. В. Кондратьевой*
Оформление серии *Т. В. Иваншина*

ЛР № 071629 от 20.04.98
Издательский Дом «ФОРУМ»
101990, Москва — Центр, Колпачный пер., д. 9а
Тел./факс: (495) 625-39-27
E-mail: forum-books@mail.ru

ЛР № 070824 от 21.01.93
Издательский Дом «ИНФРА-М»
127282, Москва, Полярная ул., д. 31в
Тел.: (495) 380-05-40
Факс: (495) 363-92-12
E-mail: books@infra-m.ru
Http://www.infra-m.ru

По вопросам приобретения книг обращайтесь:
Отдел продаж «ИНФРА-М»
127282, Москва, ул. Полярная, д. 31в
Тел.: (495) 363-42-60
Факс: (495) 363-92-12
E-mail: books@infra-m.ru

Центр комплектования библиотек
119019, Москва, ул. Моховая, д. 16
(Российская государственная библиотека, кор. К)
Тел.: (495) 695-93-15