МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
КАФЕДРА МЕДИЦИНСКОЙ И БИОЛОГИЧЕСКОЙ ФИЗИКИ

А. М. Капитонов

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие

Минск БГМУ 2013

УДК 519.2(075.8)
ББК 22.17 я73
К20

Рекомендовано Научно-методическим советом университета в качестве учебно-методического пособия 20.03.2013 г., протокол № 7

Р е ц е н з е н т ы: канд. физ.-мат. наук, доц., зав. каф. физики и высшей математики Международного государственного экологического университета им. А. Д. Сахарова В. Ф. Малишевский; канд. мед. наук, доц., зав. каф. общественного здоровья и здравоохранения Белорусского государственного медицинского университета Т. П. Павлович

Капитонов, А. М.

К20 Математическая статистика : учеб.-метод. пособие / А. М. Капитонов. – Минск : БГМУ, 2013. – 108 с.

ISBN 978-985-528-887-0.

Рассматриваются основы математической статистики. Материалы последовательно раскрывают содержание основного раздела программы курса «Высшая математика» для студентов фармацевтического факультета. Теоретические положения поясняются иллюстрациями и примерами, адаптированными для студентов медицинского университета.

Предназначено для студентов 1-го курса фармацевтического факультета.

УДК 519.2(075.8)
ББК 22.17 я73

ISBN 978-985-528-887-0

ПРЕДИСЛОВИЕ

Знакомство с основами высшей математики, в частности математической статистики, является необходимой составляющей профессиональной подготовки фармацевта. В процессе изучения высшей математики формируется способность давать точные количественные оценки на основе строгих аналитических моделей. Эта способность лежит в основе академических, профессиональных и социально-личностных компетенций, требующихся специалисту-фармацевту. В частности, владение соответствующим математическим аппаратом — залог успеха в дальнейшем изучении физики, аналитической и коллоидной химии, экономики фармации, ряда других естественнонаучных и специальных дисциплин.

Данное учебно-методическое пособие предназначено для студентов фармацевтического факультета. Оно написано на основе лекций, читаемых автором студентам БГМУ. В нем раскрыты вопросы основного (3-го) раздела курса высшей математики — математической статистики.

Для удобства использования в процессе самоподготовки студентов материал издания структурирован согласно пунктам действующей учебной программы. Последовательность изложения соответствует методическим рекомендациям автора «Вопросы и задания к практическим занятиям по высшей математике» (БГМУ, 2011).

Учебно-методическое пособие написано с учетом общего уровня математической подготовки студентов медицинского вуза. Предпочтение отдано ясности, доступности и логической последовательности излагаемого материала. Приводятся решения задач, имеющих прикладное значение для фармации и смежных дисциплин. Где требуется, строгие математические выкладки и формальные доказательства заменены подходящим иллюстративным материалом. Ставилась цель облегчить процесс усвоения читателем основных понятий высшей математики, подтолкнуть к осознанию универсальности ее подходов и логических построений.

1. ЗАДАЧА МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. СТАТИСТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ВЫБОРКИ, ГИСТОГРАММА

Объективно существующая вероятность того или иного случайного события проявляется при многократном повторении случайного эксперимента. В результате таких массовых экспериментов (наблюдений), результат которых подвержен влиянию случайных факторов, исследователь получает наборы экспериментальных данных. Как правильно и математически строго интерпретировать эти данные на основе теории вероятностей? Каковы оптимальные способы сбора экспериментальных данных? Каким образом выразить в сжатом, удобном для восприятия виде наиболее существенную информацию, содержащуюся в больших массивах данных? Ответы на эти и другие важные практические вопросы призвана дать математическая статистика.

1.1. Основная задача статистики. Понятие о законе больших чисел. Метод выборки. Генеральная и выборочная совокупности. Репрезентативность выборки

Вся информация о случайной величине полностью содержится в ее законе распределения. Если этот закон известен, то можно найти любые характеристики исследуемой случайной величины и судить о наличии связей с другими величинами. Для решения таких задач используются подходы и методы теории вероятностей. Непосредственное применение этих методов возможно, когда точно известны вероятности всех значений дискретной случайной величины или плотность вероятности непрерывной случайной величины. Однако на практике законы распределения часто остаются неизвестными. Действительно, классический метод определения вероятности применим только в тех случаях, когда все элементарные события равновероятны. Статистический метод приводит к точному результату лишь при бесконечном повторе испытаний (а этот процесс может длиться бесконечно долго). На практике же распространены ситуации, когда все, что мы знаем о случайной величине, — это ограниченный набор данных, полученных в результате конечной серии наблюдений или экспериментов над данной случайной величиной. Такие данные подвержены случайному разбросу, поэтому заключенная в них информация об исследуемой величине неполная, таит в себе неопределенность*. Соответственно, и выводы, сделанные на основании таких данных, не могут быть абсо-

* Обратимся к примеру с игральной костью. Пусть в серии из десяти бросков четные числа выпали шесть раз, а нечетные — четыре раза. Можно ли на этом основании считать, что вероятность выпадения четного числа больше, чем нечетного? Составьте свои примеры, иллюстрирующие неопределенность результатов случайных экспериментов.

лютно достоверными, они носят вероятностный характер. А полученные в результате статистического анализа значения каких-либо величин заведомо являются оценочными, они лишь приблизительно соответствуют истинным значениям этих величин.

Основная задача математической статистики — по наблюдаемым данным, подверженным влиянию случайных факторов, оценить закон распределения исследуемой величины. Во многих частных случаях статистическое исследование может ограничиваться:

1) оценкой характеристик распределения случайной величины;

2) поиском соответствий между распределениями и теоретическими моделями;

3) выявлением зависимостей между случайными величинами.

Подходы к решению задач математической статистики основаны на двух эмпирических (известных из практики) фактах: при многократном повторении эксперимента относительная частота $\frac{m}{n}$ случайного события группируется вокруг вероятности этого события p, а среднее значение случайной величины $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$ — вокруг ее математического ожидания μ. Теоретическое обоснование этих эмпирических фактов содержится в группе теорем, объединенных общим названием *закон больших чисел*. Ценность закона больших чисел состоит в том, что он открывает путь математически строгому, основанному на теории вероятностей решению задач статистики. По сути дела, закон больших чисел позволяет использовать относительную частоту массовых событий как статистический аналог их вероятности и на этой основе оценивать характеристики распределений, в частности использовать среднее значение случайной величины как статистическую оценку ее математического ожидания.

Если повторять несколько раз одну и ту же серию экспериментов по определению вероятности p случайного события, то результат (наблюдаемая относительная частота $\frac{m}{n}$) каждый раз может быть разным. Более того, величина отклонения статистической оценки $\frac{m}{n}$ от истинного значения p в любой, какой угодно длительной серии экспериментов может превысить допустимый уровень ε (греческая буква «эпсилон»). Можно, например, представить себе, что в серии бросков игральной кости выпадают исключительно четные числа. Такой результат ($\frac{m}{n} = 1$ при p = 0,5)

нельзя исключить, хотя его вероятность P становится исчезающе малой при большом числе бросков: $\lim_{n\to\infty} P = 0$.

Более строгие формулировки закона больших чисел описывают вероятности значительных (превышающих произвольную положительную величину ε) отклонений статистических оценок от истинных значений. Так, для оценки $\frac{m}{n}$ вероятности p какого-либо случайного события A имеем:

$$P\left(\left|\frac{m}{n} - p\right| > \varepsilon\right) < \frac{p\cdot(1-p)}{n\cdot\varepsilon^2}. \qquad (1.1а)$$

А для оценки $\bar{x}$ математического ожидания μ случайной величины X, обладающей дисперсией σ^2:

$$P\left(\left|\bar{x} - \mu\right| > \varepsilon\right) < \frac{\sigma^2}{n\cdot\varepsilon^2}. \qquad (1.1б)$$

В неравенствах (1.1) P обозначает вероятность того, что по результатам n наблюдений погрешности статистических оценок $\left|\frac{m}{n} - p\right|$ или $\left|\bar{x} - \mu\right|$ превысят допустимый предел ε. Как видно, эти вероятности могут быть сделаны сколь угодно малыми путем увеличения числа наблюдений n. Следовательно, используя достаточно большие массивы экспериментальных данных, можно добиться любой желаемой надежности и точности статистических оценок — получить ключ к решению задач математической статистики.

Пример. Требуется экспериментально подтвердить, что вероятность выпадения «шестерки» при броске игральной кости равна p = $\frac{1}{6}$. Сколько раз потребуется повторить эксперимент (бросок), чтобы вероятность успеха была не менее 95 %? Допускается погрешность менее: а) ε = 0,01; б) ε = 0,1.

Решение. Выразим число наблюдений n из неравенства (1.1а). Учитывая, что $n > 1$, получаем: $n > \frac{p\cdot(1-p)}{P\cdot\varepsilon^2}$.

По условию задачи вероятность отклонения наблюдаемой относительной частоты должна быть: $P\left(\left|\frac{m}{n} - \frac{1}{6}\right| > \varepsilon\right) < 1 - 0{,}95 = 0{,}05$.

Всего потребуется наблюдений:

$$n > \frac{\frac{1}{6}\cdot\left(1-\frac{1}{6}\right)}{0{,}05\cdot\varepsilon^2} = \frac{1\cdot 5}{6\cdot 6\cdot 0{,}05\cdot\varepsilon^2} = \frac{1}{0{,}36\cdot\varepsilon^2}.$$

В зависимости от допустимой погрешности:

а) если ε = 0,01, то $n > \frac{1}{0{,}36 \cdot 0{,}01^2} = \frac{10\,000}{0{,}36} > 27\,777$;

б) если ε = 0,1, то $n > \frac{1}{0{,}36 \cdot 0{,}1^2} = \frac{100}{0{,}36} > 277$.

***Ответ*:** а) $n > 27\,777$; б) $n > 277$.

Замечание. Из закона больших чисел следует, что точность статистических оценок нарастает довольно медленно, пропорционально $\sqrt{n}$. Для того чтобы повысить точность оценки в 10 раз, требуется взять в 100 раз больший объем наблюдений.

Историческое развитие статистики как прикладной науки связано с исследованиями больших групп объектов, обладающих каким-либо признаком. Таким признаком может быть, например, возраст, пол, показатели состояния здоровья и экономического благополучия и т. д. отдельных индивидуумов, образующих достаточно большую группу, например, население страны. С точки зрения исследователя такой признак может рассматриваться как случайная величина X, принимающая индивидуальное значение x_i у каждого объекта. Эти индивидуальные значения до проведения исследования неизвестны, что интерпретируется как качественная однородность объектов — одинаковый закон распределения величины X среди всех объектов.

Генеральной совокупностью называют множество качественно однородных объектов. Как правило, генеральная совокупность состоит из большого, в идеале бесконечного, количества элементов. Генеральная совокупность может реально существовать (например, партии ампул или таблеток, направляемые на контроль качества), а может быть гипотетической (пациенты, которые будут использовать данный препарат, если он успешно пройдет клинические испытания). Поэтому термин *генеральная совокупность* в математической статистике применяется и к самой исследуемой случайной величине X.

Сплошное обследование всех элементов генеральной совокупности не всегда возможно или целесообразно. Так, если контроль качества медикаментов требует разрушения ампул или таблеток, то сплошное обследование приведет к уничтожению всей партии. В таких случаях ограничиваются обследованием выборки. *Выборочная совокупность* (*выборка*) — любое подмножество (часть) элементов генеральной совокупности, отобранное для проведения статистического исследования. Например, таблетки, взятые из партии для контроля качества, являются выборкой.

Число элементов выборки называют ее *объемом* (обозначим буквой n).

Используя значения $x_1, \ldots, x_n$, попавшие в выборку, можно оценить параметры всей генеральной совокупности (исследуемой случайной величины X). В этом состоит суть *метода выборки*. Безусловно, полученные

оценки не будут абсолютно точными и надежными, поскольку выборка содержит лишь часть информации о генеральной совокупности.

Наилучшие оценки обеспечивают *репрезентативные* (*представительные*) выборки. Такую выборку можно считать набором значений $x_1, x_2, \ldots, x_n$, которые исследуемая случайная величина X приняла в результате n испытаний*. На практике репрезентативность обеспечивается случайным отбором объектов в выборку.

1.2. Статистическое распределение выборки, варианты, частоты, относительные частоты. Статистический ряд, ранжированный ряд, вариационный ряд

Каким образом, имея выборку, состоящую из дискретных значений $x_1, x_2, \ldots, x_n$, оценить закон распределения всей генеральной совокупности X ? Напомним, что закон распределения дискретной случайной величины должен задавать вероятности $p_i = P(x_i)$ для всех ее возможных значений x_i.

В качестве первого шага оценим множество возможных значений генеральной совокупности. Только возможные значения с вероятностью, большей нуля, могут попасть в состав выборки $x_1, x_2, \ldots, x_n$. Это, однако, не гарантирует, что в выборке найдутся все возможные значения величины X. С другой стороны, значения выборки могут повторяться. Следовательно, наилучшей (но не обязательно полной) оценкой множества возможных значений исследуемой величины X будет совокупность *вариант* — наблюдаемых значений выборки $x_1 = x_{\min}, \ldots, x_k = x_{\max}$ (повторяющиеся значения относятся к одной и той же варианте).

Конкретные значения $x_1, x_2, \ldots, x_n$ поступают в выборку случайным образом, в результате случайных экспериментов (наблюдений, испытаний и т. д.). Восприятие человеком информации, содержащейся в таких неупорядоченных рядах данных, может быть затруднено. Несколько проще воспринимается *вариационный ряд* — последовательность всех значений выборки, записанная в порядке их возрастания**.

Число появлений варианты x_i ($i = 1, \ldots, k$) в выборке называют *частотой* m_i этой варианты. Всего имеется k частот $m_1, \ldots, m_k$, соответствующих всем k вариантам выборки $x_1, \ldots, x_k$. Просуммировав все частоты, получим число элементов выборки — ее объем n:

$$\sum_{i=1}^{k} m_i = m_1 + \ldots + m_k = n. \qquad (1.2)$$

---

* Репрезентативную выборку можно рассматривать и как набор случайных величин $X_1, X_2, \ldots, X_n$, имеющих такой же закон распределения, что и исследуемая величина X. Эта интерпретация облегчает некоторые теоретические выкладки (см. пункт 1.5).

** В медицинской литературе наблюдается неоднозначность использования математической терминологии. Так, вариационным рядом может обозначаться статистический ряд выборки и т. д. Мы постараемся придерживаться общепринятой терминологии (см., например, статью «Вариационный ряд» в Большой Советской Энциклопедии).

В соответствии с законом больших чисел следует ожидать более частого попадания в выборку более вероятных значений исследуемой случайной величины X. Значит, есть основания надеяться, что большей наблюдаемой частоте m_i варианты x_i соответствует большая объективная вероятность этого значения $p_i = P(x_i)$. Чтобы аналогия с вероятностью p_i была полной, следует использовать *относительную частоту* $\frac{m_i}{n}$, измеряемую в тех же единицах, что и p_i — в долях или процентах. Перейдя в уравнении (1.2) к относительным частотам (разделив его на n), получим аналогию с условием нормировки вероятностей:

$$\sum_{i=1}^{k} \frac{m_i}{n} = \frac{m_1}{n} + \ldots + \frac{m_k}{n} = 1. \qquad (1.3)$$

Множество всех вариант выборки x_1, …, x_k, их частот m_1, …, m_k или относительных частот $\frac{m_1}{n}$, ..., $\frac{m_k}{n}$ называется *статистическим распределением выборки*. Записав его в виде таблицы, получим *статистический ряд выборки* (табл. 1.1). Статистическое распределение выборки служит оценкой закона распределения всей генеральной совокупности.

Таблица 1.1

Статистический ряд выборки

Варианта	x_1	x_2	…	x_k	
Частота	m_1	m_2	…	m_k	$\sum m_i = n$
Относительная частота	$\frac{m_1}{n}$	$\frac{m_2}{n}$	…	$\frac{m_k}{n}$	$\sum \frac{m_i}{n} = 1$

Замечание. Использование относительных частот в статистическом распределении выборки дает более наглядную аналогию с распределением вероятностей $p_i = P(x_i)$. В этом случае уместно указывать еще и объем выборки, позволяющий оценить надежность полученных оценок (с помощью закона больших чисел).

Пример 1. Выборка состоит из следующих значений: 2, 0, 2, 3, 0, 2, 2, 3, –1, 0, 4, 2, 2, 0, 3, 4, 2, 3, 0, –1, 3, 2, 0, 0, 3, 0, 0, 2, 0, 0, 3, –1, 2, –1, 3, 0, 3, 4, 0, 2, 4, 2, 0, 2, 4, 0, –1, 2, 3, 2. (1.4)

Определить объем выборки. Построить вариационный и статистический ряды.

Решение. Выборка включает 50 элементов, поэтому ее объем: $n = 50$. Значения в порядке возрастания (вариационный ряд): –1, –1, –1, –1, –1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4.

Варианта «–1» встречается 5 раз ($m = 5$), значит, ее относительная частота $\frac{m}{n} = 0{,}1$. Варианты «0» и «2» попали в выборку по 15 раз, их отно-

сительная частота равна 0,3. Варианта «3» встречается 10 раз, значит, $\frac{m}{n} = 0{,}2$. И варианта «4» встретилась только 5 раз, значит, $\frac{m}{n} = 0{,}1$. Статистический ряд этой выборки представлен в табл. 1.2:

Таблица 1.2

Статистический ряд выборки (1.4)

x_i	–1	0	2	3	4	
m_i	5	15	15	10	5	$\sum m_i = 50$
$\frac{m_i}{n}$	0,1	0,3	0,3	0,2	0,1	$\sum \frac{m_i}{n} = 1$

В математической статистике существуют методы (см. пункты 4.2 и 6.5), использующие для вычислений не сами выборочные значения, а их места в вариационном ряду, выражаемые *рангами* этих значений. Ранг равен порядковому номеру значения в вариационном ряду, если оно не повторяется. У повторяющихся значений ранг одинаков и равен среднему арифметическому их порядковых номеров. Таблица, в которой перечислены все значения x_i и их ранги, называется *ранжированным рядом*.

Пример 2. Построить ранжированный ряд по: 1, –1, 0, 3, 0, 3, 1, 0, 7, 2.

Решение. Вариационный ряд этой выборки: –1, 0, 0, 0, 1, 1, 2, 3, 3, 7. Первым идет значение «–1», его ранг равен 1. Места со 2-го по 4-е занимают три «нуля». Их ранг одинаков и равен $\frac{2+3+4}{3} = 3$. Затем следуют две «1», их ранги дробные $\frac{5+6}{2} = 5{,}5$. Ранг «2» равен 7. Места 8-е и 9-е занимают две «3», их ранг: $\frac{8+9}{2} = 8{,}5$. И замыкает ряд значение «7» с рангом 10.

***Ответ*:**

Значение	–1	0	0	0	1	1	2	3	3	7
Ранг	1	3	3	3	5,5	5,5	7	8,5	8,5	10

1.3. Эмпирическая функция распределения

Для описания случайной величины X может использоваться ее интегральная функция распределения: $F(x) = P(X < x)$. Эта функция суммирует вероятности (интегрирует плотность вероятности, если X — непрерывная) всех значений случайной величины X, меньших текущего x. Используя статистическое распределение выборки, можно построить эмпирическую функцию распределения $F^*(x)$:

$$F^*(x) = \frac{m(x)}{n} = \sum_{x_i < x} \frac{m_i}{n}. \qquad (1.5)$$

Здесь суммируются относительные частоты всех вариант X, меньших текущего значения x ($m(x)$ — суммарная частота вариант, меньших x).

Функция $F^*(x)$ называется эмпирической, поскольку она строится на основе эмпирических (получаемых опытным путем) частот появления вариант выборки. Это подчеркивает ее отличие от «теоретической» функции распределения $F(x)$, определяемой «теоретическими» вероятностями возможных значений генеральной совокупности.

По своим свойствам эмпирическая функция распределения $F^*(x)$ напоминает «теоретическую» $F(x)$. Обе эти функции определены для любых значений x, от $-\infty$ до $+\infty$, и они нигде не убывают (рис. 1.1). Минимальное значение $\min F^*(x) = 0$ принимается, пока x не превысит минимальной варианты выборки. Максимальное же значение $\max F^*(x) = 1$ достигается при x, больших максимальной варианты.

Пример. По выборке (1.4) построить эмпирическую функцию распределения $F^*(x)$, построить ее график.

Решение. Воспользуемся статистическим рядом выборки (табл. 1.2). Минимальная варианта «–1», значит, $F^*(x \le 1) = 0$. Когда x становится больше «–1», $F^*(x)$ увеличивается на относительную частоту этой варианты: $F^*(-1 < x \le 0) = 0{,}1$.

Когда x превышает следующую варианту «0», $F^*(x)$ последовательно увеличивается на 0,3: $F^*(0 < x \le 2) = \dfrac{m_{-1}}{n} + \dfrac{m_0}{n} = \dfrac{1}{10} + \dfrac{3}{10} = 0{,}4.$

Проведя аналогичные рассуждения для оставшихся вариант «2», «3» и «4», определим $F^*(x)$ полностью, как показано в табл. 1.3. График этой функции представлен на рис. 1.1.

Таблица 1.3

Эмпирическая функция распределения по данным выборки (1.4)

x	$x \le -1$	$-1 < x \le 0$	$0 < x \le 2$	$2 < x \le 3$	$3 < x \le 4$	$x > 4$
$F^*(x)$	0	0,1	0,4	0,7	0,9	1

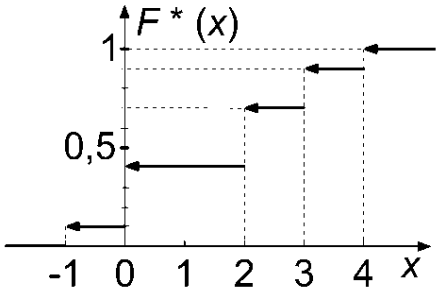

Рис. 1.1. График эмпирической функции распределения $F^*(x)$, построенной по выборке (1.4)

1.4. Графическое представление статистического распределения выборки: полигон частот и гистограмма

Для большей наглядности статистическое распределение выборки можно изобразить графически. С этой целью строятся полигоны частот и гистограммы. Полигон относительных частот выборки строится по аналогии с полигоном распределения дискретной случайной величины, а гистограмма служит статистической оценкой плотности вероятности непрерывной случайной величины.

Полигон относительных частот — график статистического распределения выборки, на котором высота точек соответствует относительным частотам $\frac{m_i}{n}$ вариант x_i, откладываемых по горизонтали. Для наглядности точки принято соединять ломаной линией. Если же строится *полигон частот*, то по вертикали откладываются частоты m_i всех вариант выборки. На рис. 1.2 в качестве примера показаны полигоны частот (a) и относительных частот (b) выборки (1.4), построенные по ее статистическому ряду (табл. 1.2).

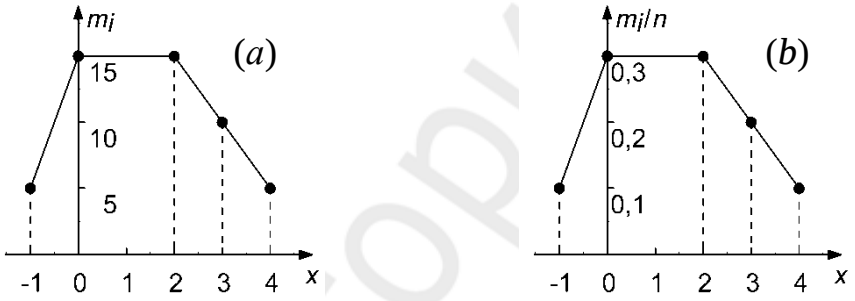

Рис. 1.2. Полигоны частот (a) и относительных частот (b) выборки (1.4), $n = 50$

Визуализация статистического распределения полигонами наиболее эффективна, если в выборку попадает относительно небольшое число вариант, а частоты вариант велики: $m_i >> 1$. Обычно такие выборки представляют дискретные случайные величины, а наблюдаемые относительные частоты $\frac{m_i}{n}$ стремятся к вероятностям p_i дискретных возможных значений, совпадающих с вариантами выборки x_i.

Если же генеральная совокупность X непрерывна, то вероятность попадания в выборку стремится к нулю у всех ее возможных значений. Вероятность же повторного отбора значения в выборку тем более нулевая. Следовательно, частоты всех вариант выборки из непрерывной генеральной совокупности равны или близки к единице ($m_i \to 1$). Полигон частот, построенный по данным такой выборки, неинформативен.

Замечание 1. На практике повторные попадания значений непрерывной случайной величины в выборку происходят, в основном, благодаря округлению результатов

измерений. Так, предположим, что показания термометра у двух пациентов равны 36,6 °С. Это отнюдь не свидетельствует о математически точном равенстве температуры тела этих двух людей. Просто разница значений в этом случае не превышает 0,1 °С.

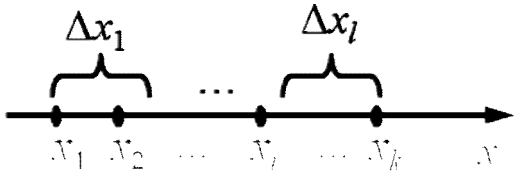

Рис. 1.3. Группировка вариант выборки $x_1, \ldots, x_k$ по l частичным интервалам

Вспомним, что закон распределения непрерывной случайной величины X может быть выражен ее плотностью вероятности: $f(x) = \frac{dP}{dx}$. Здесь dP — вероятность попадания X в бесконечно малый интервал dx. Поступим по аналогии. Сгруппируем близкие варианты выборки, как показано на рис. 1.3. Разобьем весь диапазон значений выборки, от минимальной варианты $x_1 = x_{\min}$ до максимальной $x_k = x_{\max}$ на l интервалов шириной $\Delta x_1, \Delta x_2, \ldots, \Delta x_l$. Возьмем, для простоты, ширину всех интервалов одинаковой:

$$\Delta x = \frac{x_{\max} - x_{\min}}{l}. \qquad (1.6)$$

Положение интервалов характеризуют их границы: $h_0 = x_{\min}, \ldots, h_{i+1} = h_i + \Delta x, \ldots, h_l = x_{\max}$. Найдем суммарные *частоты интервалов* $m_1^*, m_2^*, \ldots, m_l^*$. Каждая такая частота m_i^* равна числу значений, попавших в интервал $h_{i-1} \le x < h_i$. Тогда относительные частоты интервалов $\frac{m_i^*}{n}$ будут оценками теоретических вероятностей $\Delta P_i = P(h_{i-1} \le X < h_i)$ попадания величины X в эти интервалы. (Такие оценки точны при $m_i^* >> 1$, поэтому на практике, если выборка небольшая, то расхождения между $\frac{m_i^*}{n}$ и ΔP_i могут быть существенными.) Результаты группировки сведем в *интервальный* ряд (табл. 1.4). Условия нормировки (1.2) и (1.3) для интервальных частот: $\sum m_i^* = n$ и $\sum \frac{m_i^*}{n} = 1$.

Замечание 2. Мы используем интервалы $h_{i-1} \le x < h_i$ с открытыми верхними границами. Поэтому значения, попадающие точно на границу, будем относить к «правому» интервалу (к большим значениям). Значения $x_{\max}$ включаются в последний интервал $[h_{l-1}; h_l]$.

Таблица 1.4

Интервальный статистический ряд выборки, сгруппированной по *l* интервалам

Интервал	$[h_0; h_1)$	$[h_1; h_2)$	…	$[h_{l-1}; h_l]$	
Частота	m_1^*	m_2^*	…	m_l^*	$\sum m_i^* = n$
Относительная частота	$\frac{m_1^*}{n}$	$\frac{m_2^*}{n}$	…	$\frac{m_l^*}{n}$	$\sum \frac{m_i^*}{n} = 1$

Оптимальное число l интервалов группирования зависит от объема выборки n. Действительно, если интервалы Δx_i слишком узкие, то в них попадет мало значений и случайные отклонения относительных частот $\frac{m_i^*}{n}$ от вероятностей ΔP_i будут неприемлемо велики. Когда интервалы Δx_i слишком широкие, оценка закона распределения X получится грубой, поскольку суммарная частота m_i^* относится ко всему интервалу, а информация о деталях распределения X внутри интервала теряется при группировке. Считается, что оптимальное число интервалов составляет (без доказательства):

$$l \approx 1 + \log_2 n. \qquad (1.7а)$$

Перейдем от двоичного логарифма к более удобному десятичному:

$$l \approx 1 + 3{,}32 \cdot \lg n. \qquad (1.7б)$$

На практике результат вычислений по формулам (1.7) используется как приблизительный ориентир — группировка данных может проводиться и по несколько отличающемуся числу интервалов. Например, если объем выборки $n = 100$, то формула (1.7б) даст $l \approx 1 + 3{,}32 \cdot \lg 100 = 1 + 3{,}32 \cdot 2 = 7{,}64$. Такую выборку уместно группировать и по восьми, и по семи, и по девяти интервалам.

Построим график интервального ряда относительных частот. Учтем, что каждое значение $\frac{m_i^*}{n}$ характеризует свой интервал $h_{i-1} \le x < h_i$ целиком, т. е. это значение одинаково для всех точек интервала. Поэтому на графике (рис. 1.4, *а*) каждому интервалу сопоставлен прямоугольник, а не криволинейная трапеция, как на графике «теоретической» плотности вероятности $f(x)$. Такое изображение интервального ряда при помощи вертикальных прямоугольников, в основании которых лежат соответствующие интервалы $h_{i-1} \le x < h_i$, а площади равны $\frac{m_i^*}{n}$, называется *гистограмма* относительных частот.

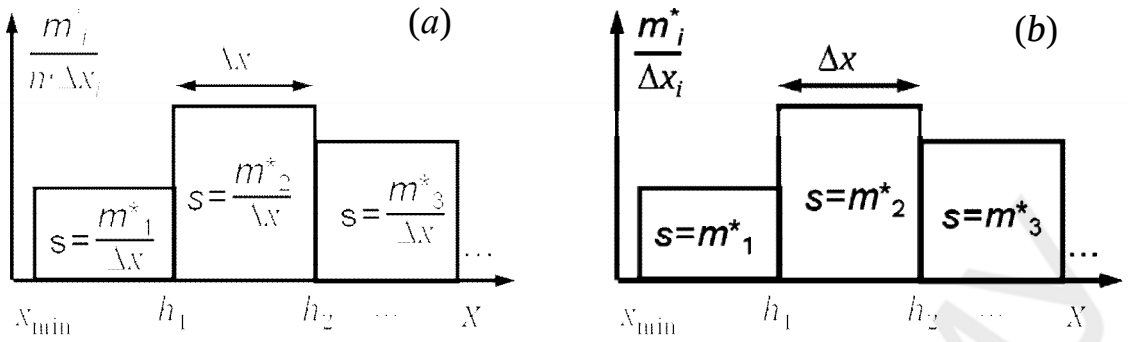

Рис. 1.4. Гистограммы относительных частот (*a*) и частот (*b*)

Важно помнить, что значение относительной частоты интервала $\frac{m_i^*}{n}$, занесенное в табл. 1.4, равно площади прямоугольника s_i, а не его высоте. Гистограмма относительных частот — статистическая оценка «теоретической» плотности вероятности $f(x)$. Поэтому высота прямоугольников на рис. 1.4, *a* равна плотности относительной частоты (отметим, что ее значения не занесены в интервальный ряд табл. 1.4):

$$f_i^* = \frac{m_i^*}{n \cdot \Delta x_i}. \qquad (1.8)$$

Ширина прямоугольников гистограммы не обязательно одинаковая. Так, интервалы с малыми частотами ($m_i^* \approx 1$) полезно присоединять к соседним. При этом суммируются не только частоты, но и ширины объединяемых интервалов.

Интервальные относительные частоты $\frac{m_i^*}{n}$ отличаются от m_i^* на постоянный множитель $\frac{1}{n}$. Поэтому гистограмма частот (рис. 1.4, *b*), построенная по тем же правилам, отличается от гистограммы на рис. 1.4, *a* лишь масштабом вертикальной оси. Высота ее прямоугольников равна плотности частот: $\frac{m_i^*}{\Delta x_i}$.

Замечание 3. В популярных компьютерных программах распространено представление данных при помощи столбиковых диаграмм, визуально похожих на гистограммы. Однако у них отображаемому значению равна высота столбика, а не площадь.

Пример. Значения непрерывной случайной величины X измерялись с округлением до целых. Оценить закон распределения X по выборке: –2, 7, –11, –1, 14, –17, 13, –10, –3, –4, –28, 1, 4, –2, 1, –2, –12, 4, 8, –4, 6, –9, –12, –21, 2, –10, –4, –4, 10, –2, 0, 2, 7, –8, –6, –2, –26, 10, –8, 3, –5, –1, –4, 11, –11, –10, –8, –9, 13, 4. (1.9)

Решение. Объем выборки $n = 50$.

Расположив значения в порядке возрастания, получим вариационный ряд: –28, –26, –21, –17, –12, –12, –11, –11, –10, –10, –10, –9, –9, –8, –8, –8, –6, –5, –4, –4, –4, –4, –4, –3, –2, –2, –2, –2, –2, –1, –1, 0, 1, 1, 2, 2, 3, 4, 4, 4, 6, 7, 7, 8, 10, 10, 11, 13, 13, 14.

Имеется $k = 27$ вариант. Диапазон значений выборки простирается от $x_{\min} = -28$ до $x_{\max} = 14$. Статистический ряд частот:

x_i	–28	–26	–21	–17	–12	–11	–10	–9	–8	–6	–5	–4	–3	–2
m_i	1	1	1	1	2	2	3	2	3	1	1	5	1	5

x_i	–1	0	1	2	3	4	6	7	8	10	11	13	14
m_i	2	1	2	2	1	3	1	2	1	2	1	2	1

Частоты малы, требуется группировка данных. Оценим оптимальное число интервалов по формуле (1.7б): $l \approx 1 + 3{,}32 \cdot \lg 50 = 1 + 3{,}32 \cdot 1{,}70 = 6{,}644$. Сгруппируем значения выборки по $l = 7$ интервалам одинаковой ширины (1.6): $\Delta x = \frac{x_{\max} - x_{\min}}{l} = \frac{14 - (-28)}{7} = 6$. Тогда границы интервалов попадут на: –28, –22, –16, –10, –4, 2, 8 и 14.

В первый интервал $x < -22$ попадают значения «–28» и «–26», поэтому $m_1^* = 2$. Во второй интервал $-22 \le x < -16$ попадают «–21» и «–17», значит, $m_2^* = 2$. В третий интервал $-16 \le x < -10$ попадают «–12» и «–11» (см. замечание 2), просуммировав их частоты, получаем $m_3^* = 4$. В четвертый интервал $-10 \le x < -4$ попали «–10», «–9», «–8», «–6» и «–5», сумма их частот равна $m_4^* = 10$. В пятом интервале $-4 \le x < 2$ встречаем «–4», «–3», «–2», «–1», «0» и «1», с учетом их частот $m_5^* = 16$. В шестой интервал $2 \le x < 8$ попали «2», «3», «4», «6» и «7», значит, $m_6^* = 9$. И в седьмом интервале $8 \le x < 14$ оставшиеся «8», «10», «11», «13» и «14», значит, $m_7^* = 6$. Вычислив $\frac{m_i^*}{50}$, получаем интервальный ряд:

Таблица 1.5

Интервальный ряд выборки (1.9), сгруппированной по семи интервалам

Интервал	[–28; –22)	[–22; –16)	[–16; –10)	[–10; –4)	[–4; 2)	[2; 8)	[8; 14)	
m_i^*	2	2	4	10	16	9	7	$\sum m_i^* = 50$
$\frac{m_i^*}{n}$	0,04	0,04	0,08	0,2	0,32	0,18	0,14	$\sum \frac{m_i^*}{n} = 1$

Чтобы построить гистограмму относительных частот, необходимо найти высоту прямоугольников по формуле (1.8). У нас ширина всех интервалов одинакова, значения f_i^* получим, поделив $\frac{m_i^*}{n}$ на $\Delta x = 6$:

Интервал	[–28; –22)	[–22; –16)	[–16; –10)	[–10; –4)	[–4; 2)	[2; 8)	[8; 14)
f_i^*	0,0067	0,0067	0,01333	0,0333	0,0533	0,03	0,0233

Замечание 4. Гистограмма — довольно грубая оценка плотности вероятности. Чтобы получить представление о детальном ходе графика $f(x)$, следует одновременно увеличивать и число интервалов ($l \to \infty$), и интервальные частоты ($m_i^* >> 1$). А это требует большого объема выборки n (хотя бы несколько сотен). Если же объем выборки ограничен, бывает полезно перестроить гистограмму несколько раз, каждый раз проводя группировку по-новому. Особенности гистограммы, сохраняющиеся при различных группировках, вероятно, присущи и $f(x)$.

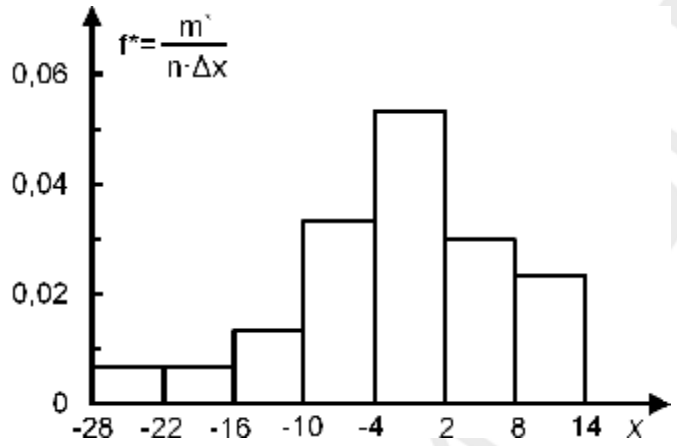

Рис. 1.5. Гистограмма относительных частот выборки (1.9), полученная группировкой ее значений по семи интервалам равной длины

1.5. Параметры статистического распределения выборки, точечные оценки характеристик генеральной совокупности, понятие о несмещенности, состоятельности и эффективности этих оценок

Важный класс задач математической статистики — оценка характеристик (параметров) распределения генеральной совокупности. При определенных обстоятельствах закон распределения считается известным. Например, когда исследуемая величина X удовлетворяет условиям центральной предельной теоремы, она распределена нормально. А нормальный закон $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ зависит только от двух параметров случайной величины X — ее математического ожидания $\mu(X)$ и среднеквадратичного отклонения $\sigma(X)$. Бывают ситуации, когда приходится довольствоваться выборками малого объема из-за сложности, длительности или дороговизны экспериментов. Имея такую выборку (когда n ограничено несколькими десятками), гораздо проще оценить отдельные, наиболее важные параметры закона распределения, чем все его детали. Наконец, существуют прикладные задачи, для решения которых знание закона рас-

пределения и вовсе не требуется, достаточно информации об основных характеристиках исследуемой генеральной совокупности.

Чтобы охарактеризовать распределения случайных величин наиболее полно, в теории вероятностей используется целый ряд параметров. Основные из них — характеристики положения (математическое ожидание, мода, медиана) и характеристики разброса (дисперсия, среднеквадратичное отклонение). Также используются характеристики формы распределения (асимметрия, эксцесс). Любой из этих параметров является функцией распределения, т. е. принимает вполне определенное численное значение Θ (греческая буква «тета») у каждой конкретной генеральной совокупности *X*. В рамках математической статистики требуется оценить это значение, используя лишь данные, попавшие в выборку. Статистическая оценка называется *точечной*, если она равна числу Θ^* (изображается точкой на числовой оси). Очевидно, что точечная оценка является функцией выборки.

Одному и тому же параметру генеральной совокупности можно давать различные оценки, воспроизводящие истинное значение с разной степенью правдоподобия. Например, в той или иной мере оценить математическое ожидание μ исследуемой величины можно, взяв среднее арифметическое выборки $\bar{x}$ или середину диапазона ее значений $\frac{x_{\max} - x_{\min}}{2}$. Какую из них использовать? Считается, что наилучшая из всех возможных точечных оценок должна быть состоятельной, несмещенной и эффективной.

При увеличении объема выборки n значения *состоятельной* оценки Θ_n^* должны группироваться вокруг оцениваемого параметра Θ. Говоря более точно, вероятность того, что погрешность состоятельной оценки $\left|\Theta_n^* - \Theta\right|$ превысит допустимый предел ε, может быть сделана сколь угодно малой путем увеличения объема выборки n:

$$\lim_{n\to\infty} P\left(\left|\Theta_n^* - \Theta\right| > \varepsilon\right) = 0. \qquad (1.10)$$

Если в формулах для вычисления характеристик теоретического распределения вероятность p заменить относительной частотой $\frac{m}{n}$, то, согласно закону больших чисел, получатся состоятельные статистические оценки этих характеристик. Данные оценки по аналогии называются выборочной дисперсией, выборочной медианой и т. д. Они используются в качестве параметров статистического распределения выборки.

Несмещенная оценка не должна содержать систематической ошибки ни при каком объеме выборки n. Другими словами, ее математическое ожидание $\mu(\Theta_n^*)$ равно оцениваемому параметру Θ при любом n:

$$\mu(\Theta_n^*) = \Theta. \qquad (1.11)$$

Статистическая оценка Θ_n^* зависит от значений выборки, отобранных случайным образом. Поэтому Θ_n^* сама может рассматриваться как случайная величина со своими собственными параметрами и законом распределения.

Наиболее *эффективная* оценка среди состоятельных и несмещенных оценок должна обладать минимальной дисперсией $D(\Theta_n^*)$. По сравнению с другими, менее эффективными оценками разброс ее значений, вычисленных по разным выборкам объема n, будет наименьшим. Эффективность оценки зависит от закона распределения генеральной совокупности X. Поэтому при решении прикладных задач требование эффективности оценок выполняется не всегда.

В качестве статистической оценки математического ожидания $\mu(X)$ генеральной совокупности X (генерального среднего) принято использовать *выборочное среднее*:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}. \qquad (1.12а)$$

Если всего в выборке k вариант x_i, частоты которых равны m_i, то:

$$\bar{x} = \frac{m_1 \cdot x_1 + m_2 \cdot x_2 + \ldots + m_k \cdot x_k}{n}. \qquad (1.12б)$$

Формула (1.12б) особенно удобна, когда в выборке присутствует небольшое количество вариант, а вычисления проводятся «вручную» по уже готовому статистическому ряду. Вычислим среднее значение выборки (1.4), используя данные из табл. 1.2:

$$\bar{x} = \frac{5 \cdot (-1) + 15 \cdot 0 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{50} = \frac{75}{50} = 1{,}5.$$

Выборочное среднее $\bar{x}$ — состоятельная оценка математического ожидания генеральной совокупности $\mu(X)$. В этом легко убедиться, сравнив формулу (1.12б) с определением математического ожидания. Если выборка велика ($n \to \infty$), то в нее должны попасть все возможные значения случайной величины x_i, а их относительные частоты $\frac{m_i}{n}$ приблизятся к вероятностям p_i.

Покажем, что $\bar{x}$ — несмещенная оценка $\mu(X)$. Представим себе гипотетическое множество всевозможных репрезентативных выборок объема n. Выборочное среднее можно рассматривать как случайную величину $\overline{X}$, возможные значения которой соответствуют различным выборкам. В таком случае и элементы выборки (первый, второй, …, n-ный) можно рассматривать как случайные величины (X_1, X_2, …, X_n) с таким же законом распределения, что и генеральная совокупность X (поскольку выборка репрезентативна).

Тогда среднее выборочное:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}. \qquad (1.12в)$$

Найдем его математическое ожидание. Учтем линейность математического ожидания $\mu(C \cdot X) = C \cdot \mu(X)$ и $\mu(X \pm Y) = \mu(X) + \mu(Y)$:

$$\mu(\overline{X}) = \mu\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{\mu(X_1) + \mu(X_2) + ... + \mu(X_n)}{n} =$$

$$= \frac{n \cdot \mu(X)}{n} = \mu(X). \qquad (1.12г)$$

Значит, условие несмещенности оценки (1.11) выполняется:

$$\mu(\overline{X}) = \mu(X). \qquad (1.13)$$

Найдем дисперсию выборочного среднего, учитывая общие свойства дисперсии: $D(C \cdot X) = C^2 \cdot D(X)$ и $D(X \pm Y) = D(X) + D(Y)$, а также то, что дисперсии элементов выборки равны дисперсии генеральный совокупности:

$$D\left(\overline{X}\right) = D\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{D(X_1) + D(X_2) + ... + D(X_n)}{n^2} =$$

$$= \frac{n \cdot D(X)}{n^2} = \frac{D(X)}{n}.$$

Таким образом, дисперсия выборочного среднего в n раз меньше дисперсии генеральной совокупности:

$$D\left(\overline{X}\right) = \frac{D(X)}{n}. \qquad (1.14)$$

Тогда среднеквадратичное отклонение выборочного среднего:

$$\sigma(\overline{X}) = \frac{\sigma(X)}{\sqrt{n}}. \qquad (1.15)$$

Мода выборки — варианта с наибольшей частотой: $m(\text{Mo}^*) = \max$. Ее можно применять только для оценки моды дискретного распределения. Если же исследуемое распределение непрерывно, то его мода соответствует максимуму плотности вероятности. Оценку моды непрерывного распределения следует давать по интервальному статистическому ряду выборки или по ее гистограмме. Так, лучшей оценкой моды распределения, представленного выборкой (1.9), будет середина пятого интервала гистограммы на рис. 1.5 — $\text{Mo}^* = -1$, а не варианты «–4» и «–2», имеющие максимальные частоты $m(-4) = m(-2) = 5$.

Медиана выборки — значение, делящее вариационный ряд выборки пополам. По аналогии с определением медианы распределения $F(\text{Me}) = 0{,}5$ для эмпирической функции распределения выполняется: $F_n^*(\text{Me}^*) = 0{,}5$.

Когда выборка содержит нечетное число значений n, ее медианой будет значение под номером $\frac{n+1}{2}$ в вариационном ряду. Если же n четное, то можно взять арифметическое среднее значений с номерами $\frac{n}{2}$ и $\frac{n}{2+1}$. Например, объем выборки (1.9) n = 50. Поэтому ее медиану определяем по 25-му и 26-му значениям вариационного ряда: $Me^* = -2$.

По аналогии с определением дисперсии в теории вероятностей $D = \mu(X - \mu)^2$ *выборочная дисперсия* равна «среднему квадрату отклонения от среднего»:

$$D^* = \overline{(x - \bar{x})^2}. \quad (1.16)$$

У выборочной дисперсии сохраняется свойство дисперсии теоретических распределений $D = \mu(X^2) - \mu^2$. Статистическими аналогами этих математических ожиданий выступают арифметические средние значения выборки (1.12) и их квадратов

$$\overline{x^2} = \frac{x_1^2 + x_2^2 + ... + x_n^2}{n} = \frac{m_1 \cdot x_1^2 + m_2 \cdot x_2^2 + ... + m_k \cdot x_k^2}{n}:$$

$$D^* = \overline{x^2} - (\bar{x})^2. \quad (1.17)$$

Это свойство может упростить некоторые теоретические выкладки, также оно полезно, когда статистическая обработка выборки проводится «вручную».

Однако выборочная дисперсия — смещенная оценка дисперсии генеральной совокупности $D(X)$. Она занижает оцениваемое значение $D(X)$ в $\frac{n}{n-1}$ раз, что особенно существенно при работе с малыми выборками.

Несмещенной оценкой дисперсии генеральной совокупности $D(X)$ является *исправленная выборочная дисперсия* S^2:

$$S^2 = \frac{n}{n-1} \cdot D^*. \quad (1.18)$$

Если исправленная выборочная дисперсия вычисляется по всем n значениям выборки $x_1, x_2, \ldots, x_n$, то формулу (1.18) можно привести к виду:

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2. \quad (1.19а)$$

Когда подсчитаны частоты $m_1, \ldots, m_k$ всех k вариант выборки $x_1, \ldots, x_k$, вычисления удобнее проводить по формуле:

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{k} m_i \cdot (x_i - \bar{x})^2. \quad (1.19б)$$

Исправленное выборочное среднеквадратичное отклонение S используется для оценки среднеквадратичного отклонения генеральной совокупности σ(*X*):

$$S = \sqrt{S^2}. \qquad (1.20)$$

Пример. Оценить среднеквадратичное отклонение случайной величины, представленной выборкой (1.4).

Решение. Используем статистический ряд этой выборки (табл. 1.2) для вычисления среднего квадрата:

$$\overline{x^2} = \frac{5\cdot(-1)^2 + 15\cdot 0^2 + 15\cdot 2^2 + 10\cdot 3^2 + 5\cdot 4^2}{50} = \frac{235}{50} = 4{,}7.$$

Выборочное среднее было вычислено ранее, $\overline{x} = 1{,}5$. Находим выборочную дисперсию по свойству (1.17): $D^* = \overline{x^2} - (\overline{x})^2 = 4{,}7 - 1{,}5^2 = 2{,}45$. Тогда исправленная выборочная дисперсия по формуле (1.18): $S^2 = \frac{n}{n-1}\cdot D^* = \frac{50}{49}\cdot 2{,}45 = 2{,}5$. А исправленное выборочное среднеквадратичное отклонение по формуле (1.20): $S = \sqrt{2{,}5} \approx 1{,}58$.

***Ответ*:** $S = \sqrt{2{,}5} \approx 1{,}58$.

2. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ. РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА. ПОГРЕШНОСТИ ИЗМЕРЕНИЙ

2.1. Метод интервальных оценок параметров генеральной совокупности. Доверительная вероятность и доверительный интервал

Отметим на числовой прямой (рис. 2.1) значение точечной оценки Θ^*, вычисленное по выборке. Истинное значение параметра Θ генеральной совокупности тоже должно изображаться на числовой прямой точкой. Однако положение точки Θ неизвестно, его мы и должны оценить. Заметим, что случайное расхождение с истинным значением Θ имеется даже у наилучшей оценки. Поэтому «растянем» точку Θ^* в интервал числовой оси. Чем шире этот интервал, тем выше вероятность накрыть им точку Θ. Суть интервальной оценки состоит в определении *доверительного интервала*, который с *доверительной вероятностью* (надежностью) γ (греческая буква «гамма») включает искомый параметр Θ исследуемой случайной величины *X*. В фармации и медицине при статистической обработке данных обычно требуется γ = 0,95 или γ = 0,99.

На рис. 2.1 изображен симметричный относительно точечной оценки Θ^* доверительный интервал. Его ширина равна 2Δ, он простирается от $\Theta^* - \Delta$ до $\Theta^* + \Delta$. Используемое значение Δ зависит от вариант, объема выборки и от требуемой доверительной вероятности γ. Когда доверитель-

ный интервал накрывает истинное значение параметра Θ, справедливо неравенство: $\left|\Theta - \Theta^*\right| \le \Delta$. (2.1)

Вероятность выполнения неравенства (2.1) и есть доверительная вероятность:

$$\gamma = \mathrm{P}(\left|\Theta^* - \Theta\right| \le \Delta). \quad (2.2)$$

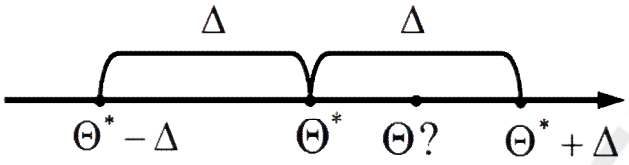

Рис. 2.1. Симметричный доверительный интервал (Θ* – Δ; Θ* + Δ) для параметра Θ

В качестве примера дадим интервальную оценку математическому ожиданию μ нормально распределенной случайной величины X с известной дисперсией σ^2. Точечной оценкой для μ(X) служит выборочное среднее $\overline{X}$. А формула доверительной вероятности (2.2) конкретизируется: $\gamma = P(\left|\mu - \bar{x}\right| \le \Delta)$. С ее помощью можно выразить Δ, если знать распределение выборочного среднего $\overline{X}$. Представим элементы выборки случайными величинами X_1, X_2, …, X_n, распределенными по одинаковому нормальному закону с плотностью вероятности $f(x) = N(x; \mu; \sigma)$. Тогда и их арифметическое среднее (1.12в) тоже будет иметь нормальное распределение. Математическое ожидание (1.12г) выборочного среднего такое же, как у генеральной совокупности, а вот среднеквадратичное отклонение (1.15) в $\sqrt{n}$ раз меньше. Следовательно, закон распределения среднего в нашем случае имеет вид:

$$f(\bar{x}) = N\left(\bar{x}; \mu; \frac{\sigma}{\sqrt{n}}\right). \quad (2.3)$$

График плотности вероятности (2.3) показан на рис. 2.2. Вероятность того, что значение $\bar{x}$ лежит между μ – Δ и μ + Δ (доверительная вероятность), равна площади заштрихованной фигуры. Ее можно выразить через функцию Лапласа (прил. 3). Вероятность попадания в симметричный интервал μ ± Δ дается формулой: $P(\mu - \varepsilon \le X < \mu + \varepsilon) = 2 \cdot \Phi\left(\frac{\varepsilon}{\sigma}\right)$. Учтем, что у $\overline{X}$ среднеквадратичное отклонение составляет $\frac{\sigma}{\sqrt{n}}$, поэтому искомая вероятность равна: $\gamma = 2 \cdot \Phi\left(\frac{\Delta \cdot \sqrt{n}}{\sigma}\right)$. Обозначим аргумент функции Лапласа z_γ:

$$z_\gamma = \frac{\Delta \cdot \sqrt{n}}{\sigma}. \quad (2.4)$$

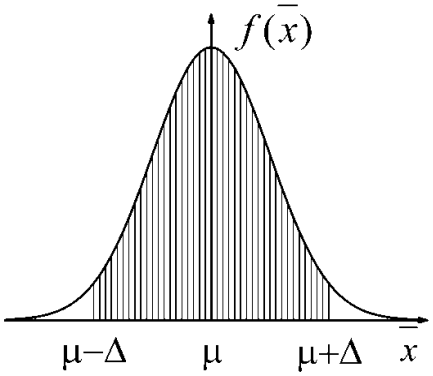

Рис. 2.2. Распределение (2.3) выборочного среднего $\overline{X}$ случайной величины *X*, подчиняющейся нормальному закону $f(x) = N(x; \mu; \sigma)$

Его можно определить (прил. 3) по известному значению функции:

$$\Phi(z_\gamma) = \frac{\gamma}{2}. \qquad (2.5)$$

Так, при $\gamma = 0{,}95$ функция Лапласа равна $\Phi(z_{0,95}) = \frac{0{,}95}{2} = 0{,}475$, а ее аргумент $z_{0,95} = 1{,}960$. Если же $\gamma = 0{,}99$, то $\Phi(z_{0,99}) = 0{,}495$ и $z_{0,99} = 2{,}576$.

Когда коэффициент z_γ известен, из формулы (2.4) можно выразить полуширину доверительного интервала Δ и записать интервальную оценку в виде $\mu = \bar{x} \pm \Delta$:

$$\mu = \bar{x} \pm z_\gamma \cdot \frac{\sigma}{\sqrt{n}}. \qquad (2.6)$$

Значение со знаком «минус» равно нижней границе доверительного интервала, а со знаком «+» — его верхней границе.

Увеличение доверительной вероятности γ требует большего z_γ и, следовательно, более широкого доверительного интервала. А чем он шире, тем менее точна интервальная оценка. Единственный способ улучшить и точность, и надежность оценки — увеличение объема выборки *n*.

Пример. Среднеквадратичное отклонение σ массы таблетки при производстве составляет 3 мг. Среднее значение $\bar{x}$, измеренное по 36 таблеткам, отобранным из партии, равно 100 мг. Дать интервальную оценку средней массы таблеток в партии. Доверительная вероятность $\gamma = 0{,}95$.

Решение. Функция Лапласа равна (2.5): $\Phi(z_\gamma) = \frac{\gamma}{2} = \frac{0{,}95}{2} = 0{,}475$. Значение ее аргумента находим в таблице значений функции Лапласа (прил. 3): $z_\gamma = 1{,}960$. Теперь имеются все данные для подстановки в формулу (2.6): $\mu = 100 \pm 1{,}960 \cdot \frac{3}{\sqrt{36}} = 100 \pm 0{,}98 \approx 100 \pm 1$.

***Ответ*:** доверительный интервал (99; 101), $\mu \approx 100 \pm 1$.

2.2. Интервальная оценка генерального среднего для нормально распределенной случайной величины с неизвестной дисперсией. Распределение Стьюдента. Число степеней свободы распределения

Точное значение σ среднеквадратичного отклонения генеральной совокупности часто неизвестно (именно так обстоит дело, когда доступная информация ограничена выборкой). В этом случае лучшее, что можно сделать, это использовать вместо σ ее статистическую оценку s — исправленное выборочное среднеквадратичное отклонение (1.20). При этом следует учесть, что s может быть меньше истинного значения σ. Поэтому доверительный интервал следует расширить, если мы хотим оставить прежней доверительную вероятность γ. Сделаем это, заменив в (2.6) коэффициент z_γ так называемым *коэффициентом Стьюдента* $t_{\gamma,\nu}$ (греческая буква ν читается «ню», $\nu = n - 1$):

$$\mu = \bar{x} \pm t_{\gamma,\nu} \cdot \frac{s}{\sqrt{n}}. \qquad (2.7)$$

Значение коэффициента Стьюдента $t_{\gamma,\nu}$ зависит не только от требуемой доверительной вероятности γ, но и от объема выборки $n = \nu + 1$. Если выборка мала, то оценка s может существенно недооценить σ. Поэтому при малых n коэффициент Стьюдента $t_{\gamma,\nu}$ превышает z_γ. А по мере увеличения выборки значение $t_{\gamma,\nu}$ стремится к z_γ, так как при $n \to \infty$ оценка s становится более точной. На практике коэффициент Стьюдента определяют по специальным таблицам (прил. 4) или с помощью компьютера.

Выразим $t_{\gamma,\nu}$ из (2.7): $t_{\gamma,\nu} = \frac{\sqrt{n}}{s} \cdot (\bar{x} - \mu)$. Значение этого выражения изменяется от выборки к выборке вместе с $\bar{x}$. Введем случайную величину:

$$T = \frac{\sqrt{n}}{s} \cdot (\overline{X} - \mu). \qquad (2.8)$$

Случайную величину, являющуюся функцией от других случайных величин, называют *статистикой*. Статистика T подчиняется *распределению Стьюдента*, плотность вероятности которого равна (справочно):

$$f_\nu(t) = A \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad (2.9)$$

где A — нормирующий множитель, а параметр ν называется числом степеней свободы.

При статистических расчетах *число степеней свободы* равно числу независимых случайных величин минус число ограничивающих условий. Статистика T зависит от n элементов выборки X_1, X_2, …, X_n, принимаю-

щих значения случайно и независимо друг от друга. На них накладывается одно ограничение — сумма всех X_i равна $n \cdot \overline{X}$, поэтому:

$$\nu = n - 1. \qquad (2.10)$$

Распределение Стьюдента симметрично относительно $t = 0$. При увеличении числа степеней свободы ν оно стремится к стандартному нормальному распределению: $f_\infty(t) = N(t; 0; 1)$. На рис. 2.3 показаны графики распределения Стьюдента с одной, пятью, тридцатью и с бесконечным числом степеней свободы.

Замечание. Каждому значению отклонения выборочного среднего $\overline{X}$ от математического ожидания μ(X) соответствует определенное значение статистики T. Пока отклонение не доходит до ±Δ (это соответствует границам доверительного интервала), Т-статистика остается внутри интервала $-t_{кр} < t < t_{кр}$ (см. рис. 4.1, *а*). Значение $t_{кр}$ выбирают таким образом, чтобы площадь под графиком распределения Стьюдента $f_\nu(t)$ в пределах от $-t_{кр}$ до $t_{кр}$ была равна требуемой доверительной вероятности γ. Эти значения $t_{кр}$ и есть коэффициенты Стьюдента (еще одно их название, двусторонние критические точки распределения Стьюдента, также используется в статистике).

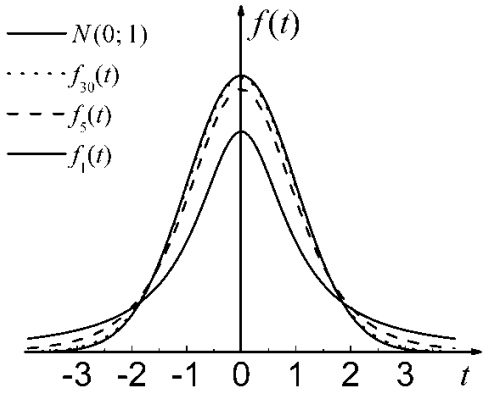

Рис. 2.3. Графики плотности вероятности $f_\nu(t)$ распределения Стьюдента (2.9) с разными числами степеней свободы (2.10). Снизу вверх ν принимает значения: 1, 5, 30 и ∞ (в последнем случае $f_\infty(t) = N(t; 0; 1)$)

Пример. Число побегов, обнаруженных на четырех растениях одного вида, составило 9, 12, 10 и 9. Дать интервальную оценку среднего числа побегов на растениях данного вида. Доверительная вероятность γ = 0,95.

Решение. Считаем число побегов нормально распределенной случайной величиной, это позволит использовать оценку (2.7).

Объем выборки $n = 4$. Число степеней свободы (2.10) ν = 4 – 1 = 3. Коэффициент Стьюдента находим в прил. 4: $t_{0,95;\,3} = 3{,}18$.

Вычислим выборочное среднее: $\bar{x} = \dfrac{9+12+10+9}{4} = 10$.

Исправленное среднеквадратичное отклонение (1.20):

$$s = \sqrt{\frac{(9-10)^2 + (12-10)^2 + (10-10)^2 + (9-10)^2}{4-1}} = \sqrt{2}.$$

Тогда интервальная оценка генерального среднего:

$$\mu = 10 \pm 3{,}18 \cdot \sqrt{\frac{2}{4}} \approx 10 \pm 2.$$

***Ответ*:** доверительный интервал (8; 12).

2.3. Абсолютная и относительная погрешности. Погрешность прямых измерений

Поговорка «точно, как в аптеке» указывает на то, что в практической работе фармацевта встречаются многочисленные измерения самых разнообразных величин. Следует понимать, что результат измерения и истинное значение измеряемой величины — это не одно и то же. Результат измерения может отклоняться от истинного значения под воздействием различных, в том числе случайных (неконтролируемых), факторов. Чтобы оценить точность измерений, принято повторять их несколько раз.

С точки зрения математики истинное значение измеряемой величины представляет собой неизвестную постоянную, обозначим ее C. А результат измерений — случайная величина X. Серия из n измерений дает выборку $x_1, x_2, \ldots, x_n$. Абсолютное значение разницы $|C - x_i|$ называется *абсолютной погрешностью* (абсолютной ошибкой) данного измерения. *Относительная погрешность* (относительная ошибка) равна отношению $\frac{|C - x_i|}{C}$, она выражает абсолютную погрешность в долях (или процентах) от истинного значения измеряемой величины.

Погрешность называется *случайной*, когда она возникает под действием случайных факторов, при этом математическое ожидание результатов измерений $\mu(X)$ равно истинному значению C: $\mu(X) = C$. *Систематические* погрешности от измерения к измерению остаются постоянными или изменяются по определенному закону, при этом $\mu(X) \neq C$. (Систематические ошибки возникают, например, при взвешивании товара не совсем честным продавцом.) Математическая статистика ограничивается анализом случайных погрешностей.

Замечание. Некоторые измерения могут дать результаты (выбросы), сильно отличающиеся от остальных. Они требуют особого внимания. Выбросы могут возникать в результате грубой ошибки (промаха). Например, показания некоторых дозиметров «зашкаливают» при измерениях вблизи высоковольтных ЛЭП. Это обусловлено электрическими помехами от ЛЭП и никак не связано с уровнем радиоактивного загрязнения таких мест. Промахи следует исключать из статистического анализа, но обязательно указывать об этом в описании проведенного исследования.

Математическая статистика ограничивается анализом влияния случайных погрешностей. Точные значения этих погрешностей неизвестны, поскольку неизвестно $C = \mu(X)$. Лучшее, что можно сделать в такой ситу-

ации, — использовать интервальную оценку математического ожидания результата измерения μ(*X*). Обычно такая оценка дается по формулам (2.6) или (2.7), справедливым, когда случайная величина *X* распределена нормально. Условие нормальности распределения *X* на практике в большинстве случаев выполняется, поскольку случайная ошибка измерения обычно складывается под воздействием большого числа случайных разнонаправленных факторов, сопоставимых по силе своего воздействия. (А в соответствии с центральной предельной теоремой распределение суммы большого числа сопоставимых случайных величин стремится к нормальному закону.) Более того, даже если распределение *X* не является нормальным, то распределение выборочного среднего $\overline{X}$ стремится к нормальному (рис. 2.2) у больших ($n \to \infty$) выборок.

В качестве точечной оценки истинного значения измеряемой величины $C = \mu(X)$ выступает среднее арифметическое результатов измерений (1.12а):

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

Чтобы оценить абсолютную погрешность $\Delta\bar{x}$ среднего, используется полуширина доверительного интервала (2.6) или (2.7). Для этого требуется значение среднеквадратичного отклонения результатов измерений. Если известно его точное значение σ, то в соответствии с формулой (2.6):

$$\Delta\bar{x} = z_\gamma \cdot \frac{\sigma}{\sqrt{n}}. \qquad (2.11)$$

Коэффициент z_γ определяют по значениям функции Лапласа (прил. 3) и требуемой доверительной вероятности по формуле (2.5): $\Phi(z_\gamma) = \frac{\gamma}{2}$.

На практике более распространена ситуация, когда доступная информация ограничена выборкой результатов измерений. В этом случае σ заменяют его выборочной оценкой *s*. А абсолютную погрешность оценивают в соответствии с формулой (2.7):

$$\Delta\bar{x} = t_{\gamma,\nu} \cdot \frac{s}{\sqrt{n}}. \qquad (2.12а)$$

Раскроем это выражение, используя формулу (1.8):

$$\Delta\bar{x} = t_{\gamma,\nu} \cdot \sqrt{\frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}. \qquad (2.12б)$$

Коэффициент Стьюдента $t_{\gamma,\nu}$ определяют (прил. 4) по требуемой доверительной вероятности γ и числу степеней свободы $\nu = n - 1$, где *n* — имеющееся количество измерений.

Исходя из оценок $\Delta\bar{x}$ для абсолютной погрешности и $\bar{x}$ для значения измеряемой величины, можно оценить относительную погрешность:

$$\delta x = \left|\frac{\Delta\bar{x}}{\bar{x}}\right|. \qquad (2.13)$$

Пример 1. По результатам n = 100 измерений вычислены среднее значение измеряемой величины $\bar{x}$ = 148 и исправленная дисперсия s^2 = 36. Определить абсолютную и относительную погрешности этих измерений с доверительной вероятностью γ = 0,95.

Решение. Большой объем выборки позволяет считать, что среднее значение распределено приблизительно нормально и оценка дисперсии s^2 близка к ее истинному значению. Поэтому абсолютную погрешность можно оценить по формуле (2.11). Коэффициент $z_{0,95}$ = 1,960 находим в прил. 3. Тогда: $\Delta\bar{x} = 1{,}960 \cdot \frac{\sqrt{36}}{\sqrt{100}} = 1{,}176.$

Оценка относительной погрешности: $\delta\bar{x} = \left|\frac{\Delta\bar{x}}{\bar{x}}\right| = \frac{1{,}176}{148} \approx 0{,}008 = 0{,}8\ \%.$

Ответ: $\Delta\bar{x}$ = 1,176; $\delta\bar{x} \approx 0{,}8$ %.

Пример 2. Результаты измерений: 47, 71, 62. Оценить истинное значение измеряемой величины и относительную погрешность полученного результата. Требуется доверительная вероятность γ = 0,95.

Решение. Объем имеющейся выборки мал: n = 3. Будем предполагать, что результаты измерений подчиняются нормальному распределению, тогда можно использовать оценку (2.12).

Среднее значение результатов измерений: $\bar{x} = \frac{47 + 71 + 62}{3} = 60.$

Коэффициент Стьюдента при γ = 0,95 и числе степеней свободы ν = 3 – 1 = 2 составляет $t_{0,95;\,2}$ = 4,30 (прил. 4). Тогда абсолютная погрешность: $\Delta x = 4{,}30 \cdot \sqrt{\frac{(47-60)^2 + (71-60)^2 + (62-60)^2}{3 \cdot (3-1)}} = 4{,}30 \cdot \sqrt{49} = 30{,}1.$

Относительная погрешность: $\delta x = \frac{30{,}1}{60} \approx 0{,}5 = 50\ \%.$

Ответ: $x \approx 60 \pm 30$; δx = 50 %.

2.4. Погрешность косвенных измерений

Встречаются ситуации, когда значение интересующей нас величины Z по тем или иным причинам нельзя измерить, но можно вычислить как функцию $Z(X, Y, \ldots)$ других величин X, Y и др., измеряемых напрямую.

Вычисленное таким путем значение величины Z называется *косвенным* измерением. Чтобы дать точечную оценку $\bar{z}$ истинному значению μ(Z) косвенно измеряемой величины, можно использовать выборочные средние ($\bar{x}, \bar{y}, \ldots$) его аргументов:

$$\bar{z} = Z(\bar{x}, \bar{y}, ...). \qquad (2.14)$$

Погрешность $\Delta\bar{z}$ зависит не только от погрешностей аргументов $\Delta\bar{x}$, $\Delta\bar{y}$ т. д., но и от вида функции $Z(X, Y, \ldots)$. Рассмотрим случай, когда $Z(X)$ вычисляется по одной непосредственно измеряемой величине X. Если погрешности $\Delta\bar{x}$ и $\Delta\bar{z}$ малы, то они связаны как дифференциалы функции и аргумента:

$$\Delta\bar{z} = z'_x \cdot \Delta\bar{x}. \qquad (2.15)$$

Численное значение производной z'_x находят, подставляя в формулу (полученную дифференцированием $Z(X)$) среднее значение аргумента $\bar{x}$.

Пример 1. Дать интервальную оценку объему куба $V = x^3$, если измеренная длина его ребра составляет: $\bar{x} = 10 \pm 1$.

Решение. Точечная оценка объема куба: $\bar{V} = (\bar{x})^3 = 1000.$ С учетом того, что $(x^3)' = 3x^2$, погрешность (2.15): $\Delta\bar{V} = 3 \cdot (\bar{x})^2 \cdot \Delta x = 3 \cdot 10^2 \cdot 1 = 300.$

***Ответ*:** $V = 1000 \pm 300$.

Перейдем к случаю, когда $Z(X, Y, \ldots)$ зависит от нескольких аргументов, измеряемых независимо друг от друга. Каждый из них независимо от других аргументов дает вклад вида (2.15) в дисперсию точечной оценки $\bar{Z} = Z(\bar{X}, \bar{Y}, ...)$. Эти вклады суммируются квадратично:

$$\Delta\bar{z} = \sqrt{(z'_x \cdot \Delta\bar{x})^2 + (z'_y \cdot \Delta\bar{y})^2 + ...} \qquad (2.16)$$

В формуле (2.16) используются частные производные z_x', z_y' и т. д.

Пример 2. Оценить объем цилиндра $V = \pi \cdot h \cdot r^2$, если результаты измерений его высоты и радиуса: $h = 10 \pm 1$, $r = 5 \pm 1$.

Решение. Точечная оценка объема: $\bar{V} = \pi \cdot \bar{h} \cdot (\bar{r})^2 = \pi \cdot 10 \cdot 5^2 = 250\pi.$

Чтобы определить погрешность по формуле (2.16), находим частные производные: $(V)_h' = (\pi \cdot h \cdot r^2)_h' = \pi \cdot (h)_h' \cdot r^2 = \pi \cdot r^2$ и $(V)_r' = (\pi \cdot h \cdot r^2)_r' = \pi \cdot h \cdot (r^2)_r' = 2\pi \cdot h \cdot r$.

Тогда погрешность объема цилиндра:

$$\Delta\bar{V} = \sqrt{(\pi \cdot 5^2 \cdot 1)^2 + (2\pi \cdot 10 \cdot 5 \cdot 1)^2} = 25\pi\sqrt{17}.$$

***Ответ*:** $V = 250\pi + 25\pi\sqrt{17}$.

3. СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ, КРИТЕРИИ ИХ ПРОВЕРКИ

3.1. Нулевая и альтернативная статистические гипотезы. Ошибки первого и второго рода

Типичный вопрос, на который люди пытаются найти ответ, находясь в ситуации недостатка информации: «Может ли такое быть?» Обычно требуется выбрать лишь один из двух взаимоисключающих ответов, «Да» или «Нет». Оба варианта являются гипотезами, нуждающимися в обосновании. Поскольку выбор делается в условиях неопределенности, на основе неполной информации, ответ может оказаться неверным. Поэтому предпочтение отдается той гипотезе, ошибочный выбор которой приведет к менее тяжелым последствиям.

В математической статистике метод гипотез применяется, когда надо установить характер неизвестного распределения, сравнить распределения или их параметры. *Статистическая гипотеза* — это предположение о неизвестном законе распределения или о параметре известного распределения. Примеры статистических гипотез: «Распределение величины X нормальное», «Дисперсии генеральных совокупностей X и Y равны между собой» и т. д. Метод статистических гипотез обычно применяют для оценки существенности различий, наблюдаемых в выборках случайных величин. Эти различия могут быть просто случайными. Например, если среднее значение в серии измерений оказалось равно 63, то может ли истинное значение измеряемой величины быть равным 50? Случайно наблюдаемое различие (между 63 и 50) или нет?

Задачи такого типа требуют формализации, перевода их на язык математики. Это делают, выдвигая пару несовместных гипотез, отражающих условия задачи:

1. *Нулевая* (*основная*) *гипотеза* H_0 предполагает, что различия несущественны, случайны. Эта гипотеза считается основной, так как именно она подвергается статистической проверке.

2. *Альтернативная* (*конкурирующая*) *гипотеза* H_1 предполагает, что различия неслучайны и существенны.

Из этой пары гипотез справедлива лишь одна. Выбрав какую-либо из них, мы отвергаем другую. Наш выбор, основанный на данных случайной выборки, может быть ошибочным. Возможны два рода ошибок:

1. *Ошибка 1-го рода* — при верной H_0 принимаем ложную H_1 (случайное различие считаем существенным).

2. *Ошибка 2-го рода* — принимаем ложную H_0 (ошибочно пренебрегаем существенным различием).

3.2. Критерии проверки статистических гипотез, законы распределения критериев, критические точки

Какой же из двух гипотез отдать предпочтение, H_0 или H_1? Правило, в соответствии с которым принимают или отвергают статистическую гипотезу, называется *критерием*. Используя критерий, оценивают соответствие гипотезы опытным данным (выборке). Для этого делают допущение, что гипотеза верна. Находят, какова при этом допущении вероятность получить именно те данные, что наблюдаются в выборке. (Например, какова вероятность получить выборку со средним значением 63, если бы математическое ожидание исследуемой величины было равно 50?) Нулевая гипотеза принимается, когда эта вероятность не меньше требуемой доверительной вероятности.

Чтобы оценить вероятность таких значений, которые наблюдаются в выборке, вычисляют некоторую функцию выборки $K(X_1, X_2, \ldots, X_n)$, называемую *статистикой критерия*. С учетом того, что статистика K зависит от случайных элементов выборки $X_1, X_2, \ldots, X_n$, ее саму можно отнести к случайным величинам. Поэтому статистика $K(X_1, X_2, \ldots, X_n)$ должна иметь свой собственный закон распределения. Статистики различных критериев могут отличаться.

Статистика K, по сути, служит мерой различий, наблюдаемых на опыте. Весь диапазон ее возможных значений делится на две части (рис. 3.1):

1) область принятия основной гипотезы H_0. Если найденное по выборке значение статистики попадает в нее, то различия признаем случайными;

2) «критическая» область — если значение K попадает в нее, то принимаем альтернативную гипотезу H_1, а H_0 считаем не соответствующей результатам наблюдений. В этом случае различия считаем существенными. Критическая область может быть односторонней (рис. 3.1, a) или двусторонней (рис. 3.1, b).

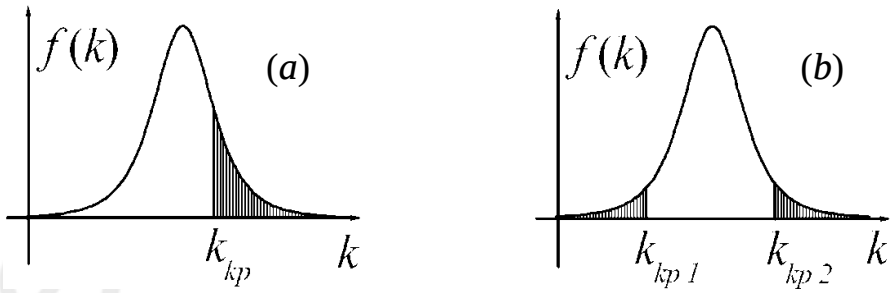

Рис. 3.1. Распределение $f(k)$ статистики критерия K. Заштрихованы участки, соответствующие критическим областям:
a — правосторонней $k > k_{кр}$; b — двусторонней, объединяющей участки $k < k_{кр1}$ и $k > k_{кр2}$

Критические точки $k_{кр}$ отделяют критическую область от области принятия основной гипотезы H_0.

3.3. Уровень значимости и мощность критериев

Даже при справедливой основной гипотезе H_0 наблюдаемые различия могут случайно оказаться настолько большими, что значение статистики K попадет в критическую область. В этом случае критерий совершит ошибку 1-го рода. На рис. 3.1 вероятность этого события равна площади заштрихованных участков распределения. Вероятность совершить ошибку 1-го рода получила название *уровень значимости* (обозначается α). Он связан с доверительной вероятностью γ соотношением:

$$\alpha = 1 - \gamma. \qquad (3.1)$$

В то же время, наблюдаемые на опыте различия могут случайно оказаться малыми, даже когда на самом деле справедлива альтернативная гипотеза H_1. Это приведет к ошибке 2-го рода, вероятность которой обозначим β. В статистике принято использовать *мощность критерия*, равную 1 – β (т. е. мощность критерия равна вероятности не совершить ошибку 2-го рода).

Статистический критерий тем лучше, чем выше его мощность 1 – β при наперед заданном уровне значимости α. Однако и здесь проявляется закон больших чисел: выбрав меньшее α (например, α = 0,01 вместо α = 0,05), мы получим снижение 1 – β. Одновременно улучшить и уровень значимости, и мощность используемого критерия можно, увеличив число наблюдений (объем выборки).

3.4. Z-критерий

Z-критерий используется для проверки равенства математических ожиданий μ(X) и μ(Y) двух нормально распределенных случайных величин с известными дисперсиями σ_X^2 и σ_Y^2. Когда объемы выборок n_X и n_Y велики, Z-критерий становится применим, даже если распределения исследуемых величин отличаются от нормального. (Поскольку в пределе $n_X \to \infty$ и $n_Y \to \infty$ распределения выборочных средних $\overline{X}$ и $\overline{Y}$ стремятся к нормальному, а оценки дисперсий — к их истинным значениям: $s_X^2 \to \sigma_X^2$ и $s_Y^2 \to \sigma_Y^2$.)

Применим Z-критерий к частному случаю. Проверим равенство математического ожидания μ(X) нормально распределенной случайной величины наперед заданному значению m. Сформулируем основную гипотезу H_0: $\mu = m$. В качестве альтернативной выберем гипотезу H_1: $\mu \neq m$. Основная гипотеза согласуется с данными наблюдений (выборкой), если предполагаемое значение m попадает в доверительный интервал (2.6): $\mu = \bar{x} \pm z_\gamma \cdot \frac{\sigma}{\sqrt{n}}$. Мерой допустимого (случайного) различия здесь является $z_\gamma = (\bar{x} - \mu) \cdot \frac{\sqrt{n}}{\sigma}$.

Введем по аналогии *Z-статистику*:

$$Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}. \qquad (3.2)$$

Z-статистика (3.2) служит мерой отклонения выборочного среднего $\overline{X}$ от предполагаемого значения m. Поскольку $\overline{X}$ распределено нормально, а в определении (3.2) использована стандартизирующая замена для $\overline{X}$, Z-статистика подчиняется стандартному нормальному распределению $f(z) = N(z; 0; 1)$. Поскольку случайные отклонения выборочного среднего $\overline{X}$ в сторону значений, больших чем $\mu(X)$, имеют такую же вероятность, что и отклонения в сторону меньших значений, область принятия H_0 симметрична относительно $z = 0$, как показано на рис. 3.2, *а*. Критические точки $\pm z_{кр}$ соответствуют попаданию m на границы доверительного интервала. Они отсекают критическую область, состоящую из двух равновеликих «хвостов». Суммарная площадь этих «хвостов» должна быть равна требуемому уровню значимости α. А площадь каждого из них $\frac{\alpha}{2}$ в сумме со значением функции Лапласа $\Phi(z_{кр})$ дает $\frac{1}{2}$ (половину всей площади распределения). Поэтому *двустороннее* критическое значение Z-статистики может быть найдено из:

$$\Phi(z_{кр}) = \frac{1}{2} - \frac{\alpha}{2}. \qquad (3.3)$$

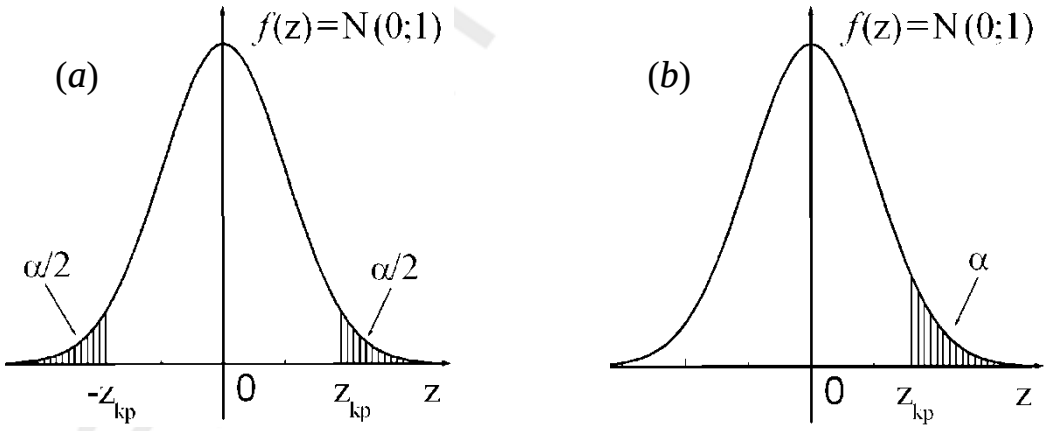

Рис. 3.2. График распределения Z-статистики: $f(z) = N(z; 0; 1)$. Заштрихованы участки, соответствующие критической области двусторонней (*а*); правосторонней (*b*)

При уровне значимости α = 0,05 двусторонняя критическая точка $z_{кр} = 1{,}960$, а при α = 0,01 будет $z_{кр} = 2{,}576$.

Пример 1. При измерениях величины X данным прибором $\sigma = 2$. По результатам $n = 25$ измерений в пробе получено среднее значение $\bar{x} = 19$. Можно ли считать, что на уровне значимости $\alpha = 0{,}01$ подтверждено отклонение от стандарта, равного $m = 20$?

Решение. Основная гипотеза H_0: $\mu = 20$; альтернативная гипотеза H_1: $\mu \neq 20$.

Считаем распределение результатов измерений нормальным. Это позволяет использовать Z-критерий. Вычислим Z-статистику (3.2): $z = (19 - 20) \cdot \dfrac{\sqrt{25}}{2} = 2{,}5$.

Двустороннее критическое значение Z-статистики на требуемом уровне значимости определим в соответствии с формулой (3.3): $\Phi(z_{кр}) = \dfrac{1}{2} - \dfrac{0{,}01}{2} = 0{,}495$ и по таблице значений функции Лапласа (прил. 3) находим $z_{кр} = 2{,}576$.

$|z| < z_{кр}$, значит, различие не доказано на уровне значимости $\alpha = 0{,}01$.

***Ответ*:** нет. (При $\alpha = 0{,}01$ нет оснований считать, что $\mu \neq m$.)

Замечание. При наличии компьютера или достаточно подробных таблиц вероятность ошибки 1-го рода, соответствующую значению Z-статистики, можно определить точно. Так, для $z = 2{,}5$ получаем $\alpha = 0{,}0062$.

Если достоверно известно, что математическое ожидание μ исследуемой генеральной совокупности не может быть больше (или не может быть меньше), чем предполагаемое значение m, то альтернативная гипотеза H_1 будет односторонней. Правосторонняя H_1 предполагает, что $\mu > m$, а левосторонняя H_1: $\mu < m$. Остановимся, для определенности, на правосторонней H_1: $\mu > m$. В ее пользу говорят только выборки, чье среднее значение $\bar{x}$ превышает m. Для них Z-статистика всегда положительна. В таком случае критическая область будет правосторонней, как показано на рис. 3.2, *b*. Поэтому требуемый уровень значимости α равен площади одного «хвоста» распределения, и *одностороннее* критическое значение Z-статистики можно найти из условия:

$$\Phi(z_{кр}) = \frac{1}{2} - \alpha. \qquad (3.4)$$

При уровне значимости $\alpha = 0{,}05$ односторонняя критическая точка $z_{кр} = 1{,}645$, а при $\alpha = 0{,}01$ получается $z_{кр} = 2{,}326$. Односторонние критические значения меньше, чем двусторонние. Это обусловлено тем, что односторонняя альтернативная гипотеза требует более полной информации об исследуемой случайной величине.

Пример 2. При измерениях величины X данным прибором $\sigma = 2$. По результатам $n = 25$ измерений в пробе получено среднее значение $\bar{x} = 19$. Можно ли считать, что истинное значение $\mu(X)$ меньше стандарта, равного $m = 20$? Требуемый уровень значимости $\alpha = 0{,}01$.

Решение. Основная гипотеза H_0: $\mu = 20$; альтернативная гипотеза H_1: $\mu < m$.

Значение Z-статистики берем из предыдущего примера: $z = (19 - 20) \cdot \frac{\sqrt{25}}{2} = 2{,}5$.

Значение функции Лапласа для односторонней H_1 и $\alpha = 0{,}01$ находим из условия (3.4): $\Phi(z_{кр}) = \frac{1}{2} - 0{,}01 = 0{,}49$. Ему соответствует односторонняя критическая точка $z_{кр} = 2{,}326$.

$|z| > z_{кр}$, значит, на уровне значимости $\alpha = 0{,}01$ принимаем H_1: $\mu < m$.

***Ответ*:** да. (При $\alpha = 0{,}01$ есть основания считать, что $\mu < m$.)

Перейдем к проверке равенства истинных средних $\mu(X)$ и $\mu(Y)$ двух случайных величин. Рассмотрим случайную величину $X - Y$. Ее математическое ожидание: $\mu(X - Y) = \mu(X) - \mu(Y)$. Значит, основную гипотезу $\mu(X) = \mu(Y)$ можно заменить на H_0: $\mu(X - Y) = 0$.

Изменим определение Z-статистики (3.2) так, чтобы оно отвечало новой формулировке основной гипотезы H_0: $\mu(X - Y) = 0$. Поскольку теперь исследуется случайная величина $X - Y$ и предполагаемое $m = 0$, то в числителе $(\bar{X} - m)$ следует заменить на $(\bar{X} - \bar{Y})$. Знаменатель (3.2) был равен среднеквадратичному отклонению числителя: $\frac{\sigma}{\sqrt{n}} = \sigma(\bar{X}) = \sigma(\bar{X} - m)$. Теперь в знаменателе будет $\sigma(\bar{X} - \bar{Y})$. Используем свойство дисперсии: $\sigma^2(\bar{X} \pm \bar{Y}) = \sigma^2(\bar{X}) + \sigma^2(\bar{Y})$ и дисперсии средних $\sigma^2(\bar{X}) = \frac{\sigma_X^2}{n_X}$ и $\sigma^2(\bar{Y}) = \frac{\sigma_Y^2}{n_Y}$.

Тогда в знаменателе будет $\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$, а Z-статистика примет вид:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}. \qquad (3.5)$$

Пример 3. Выборка X включает $n_X = 100$ представителей одной возрастной группы, а выборка Y — $n_Y = 150$ другой. Средние значения биохимического показателя в выборках составляют $\bar{X} = 125$ и $\bar{Y} = 115$, а

оценки среднеквадратичных отклонений $\sigma_X = 20$ и $\sigma_Y = 18$. Свидетельствуют ли эти данные об изменении данного показателя с возрастом? Использовать $\alpha = 0{,}05$.

Решение. Основная гипотеза H_0: $\mu(X) = \mu(Y)$; альтернативная гипотеза H_1: $\mu(X) \neq \mu(Y)$ — двусторонняя.

Z-критерий применим, так как выборки велики. Вычислим Z-статистику (3.5):

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} = \frac{125 - 115}{\sqrt{\frac{20^2}{100} + \frac{18^2}{150}}} = \frac{10}{\sqrt{4 + 2{,}025}} \approx 4{,}03.$$

Значение функции Лапласа для двусторонней H_1 и $\alpha = 0{,}05$ по условию (3.4): $\Phi(z_{кр}) = \frac{1}{2} - \frac{0{,}05}{2} = 0{,}475$. Ему соответствует критическое значение $z_{кр} = 1{,}960$.

$|z| > z_{кр}$, значит, на $\alpha = 0{,}01$ принимаем H_1: $\mu(X) \neq \mu(Y)$.

Ответ: да.

3.5. Непараметрические критерии проверки статистических гипотез, критерий знаков

Статистические критерии можно разделить на параметрические и непараметрические. *Параметрические* критерии используют предположения о законе распределения генеральной совокупности и позволяют оценивать его математические параметры. Предполагаемый закон распределения заимствуется из теории вероятностей и выражается математическими формулами. Такие «теоретические» распределения полностью определяются несколькими параметрами. Так, нормальное распределение полностью определено, когда известны его математическое ожидание μ и среднеквадратичное отклонение σ. Как раз такие параметры и оценивают с помощью параметрических критериев. Наиболее широкий круг практического применения нашли параметрические критерии, связанные с нормальным распределением. Рассмотренный выше Z-критерий является параметрическим, поскольку с его помощью проверяют гипотезы о математическом ожидании (параметре) нормально распределенных случайных величин.

Непараметрические критерии не используют предположений о законе распределения генеральной совокупности. В частности, непараметрическим является *критерий знаков*. Он используется для проверки возможного равенства медианы Me непрерывной генеральной совокупности X наперед заданному значению m. Критерий знаков применим к любой непрерывной случайной величине X с каким угодно распределением. («Цена» такой универсальности — переход от математического ожидания μ к медиане Me.)

Критерий знаков использует интервальную оценку медианы Me генеральной совокупности. Если предполагаемое значение m попадает в доверительный интервал, то принимается основная гипотеза H_0: Me = m.

Поскольку генеральная совокупность X непрерывна, вероятность попадания в выборку значения, точно равного предполагаемой медиане m, равно нулю. Совпадения могут появиться в результате округления. Поэтому значения, равные m, из выборки выбрасываются. Каждое оставшееся значение могло с равной вероятностью оказаться или больше, или меньше истинного значения медианы. Следовательно, при случайном отборе этих значений в репрезентативную выборку выполнялись условия схемы Бернулли: серия независимых испытаний, исходом каждого могло быть или событие x > Me, или противоположное ему x < Me, и вероятности этих событий оставались постоянными в течение всех испытаний. Поэтому количество n_+ элементов выборки, больших медианы Me генеральной совокупности, и количество n_- значений выборки, меньших Me, подчиняются биномиальному распределению $P_m = C_n^m \cdot p^m \cdot q^{n-m}$ с $p = q = \frac{1}{2}$ и числом испытаний:

$$n' = n_+ + n_-. \tag{3.6}$$

В качестве статистики критерия знаков можно использовать n_+ или n_-. Для определенности возьмем n_+. При справедливой H_0 вероятность того, что ровно n_+ значений из n' будут больше медианы, составляет:

$$p(n_+) = \left(\frac{1}{2}\right)^{n'} \cdot C_{n'}^{n_+}, \tag{3.7}$$

где $C_{n'}^{n_+} = \frac{n'!}{n_+! \cdot (n' - n_+)!} = \frac{n'!}{n_+! \cdot n_-!}$ — биномиальные коэффициенты (прил. 1).

Сомнение в справедливости нулевой гипотезы H_0: Me = m могут вызвать те выборки, у которых n_+ или слишком мало, или слишком велико. Поэтому в критическую область может попасть определенное число крайних значений выборки. При правосторонней альтернативной гипотезе H_1: Me > m в критическую область следует включать s самых больших значений выборки, пока сумма их вероятностей не перекроет требуемого уровня значимости α:

$$\left(\frac{1}{2}\right)^{n'} \cdot \sum_{i=0}^{s} C_{n'}^i \le \alpha. \tag{3.8а}$$

Эта формула справедлива и при левосторонней альтернативной гипотезе H_1: Me < m, однако теперь в критическую область попадут s самых маленьких значений выборки. При двусторонней альтернативной гипотезе H_1: Me ≠ m критическая область симметричная. Поэтому в нее включа-

ют по s крайних значений выборки, так чтобы с каждой стороны выполнялось:

$$\left(\frac{1}{2}\right)^{n'} \cdot \sum_{i=0}^{s} C_{n'}^{i} \le \frac{\alpha}{2}. \qquad (3.8б)$$

Основная гипотеза H_0: Me = m принимается, когда предполагаемое значение медианы m не попадает в критическую область.

Пример 1. Вариационный ряд выборки: –1, 0, 0, 0, 1, 1, 2, 3, 3, 7. Может ли медиана генеральной совокупности быть равна 2? Использовать α = 0,05.

Решение. Основная гипотеза H_0: Me = 2, альтернативная — H_1: Me ≠ 2. Количество значений выборки, превышающих 2: $n_+ = 4$, а всего ≠ 2: $n' = 9$.

Альтернативная гипотеза двусторонняя, поэтому критическую область определяем в соответствии с (3.8б): $\left(\frac{1}{2}\right)^{9} \cdot \sum_{i=0}^{s} C_{9}^{i} \le \frac{0{,}05}{2}$. Выражаем отсюда сумму биномиальных коэффициентов: $\sum_{i=0}^{s} C_{9}^{i} \le \frac{0{,}05 \cdot 2^{9}}{2} = 0{,}1 \cdot 2^{7} =$ = 12,8. По прил. 1 находим, что при числе испытаний $n' = 9$ сумма двух крайних биномиальных коэффициентов равна 1 + 9 = 10, а сумма трех крайних 1 + 9 + 36 = 46 уже превышает 12,8. Поэтому в критическую область попадают по $s = 2$ крайних значения. Это «–1», «0» слева и «3», «7» справа. Значит, доверительный интервал для медианы генеральной совокупности: 0 < Me < 3. Предполагаемое значение $m = 2$ в него входит, поэтому принимается H_0.

***Ответ*:** да, на уровне значимости α = 0,05 принимается H_0: Me = 2.

Согласно теореме Муавра–Лапласа, при росте числа наблюдений $n' \to \infty$ биномиальное распределение (3.7) стремится к нормальному с математическим ожиданием $\frac{n'}{2}$ и дисперсией $\frac{\sqrt{n'}}{2}$. Поэтому в случае больших n' сумма (3.8) вероятностей s крайних значений стремится к $\frac{1}{2} - \Phi\left(\frac{\frac{n'}{2} - s}{\frac{\sqrt{n'}}{2}}\right)$. Здесь критическое значение s в аргументе функции Лапласа соответствует критической точке Z-статистики $z_{кр} \approx \frac{\frac{n'}{2} - s}{\frac{\sqrt{n'}}{2}}$. При

конечном числе наблюдений n' это выражение нуждается в поправке: $z_{кр} \approx \dfrac{\frac{n'}{2} - s - \frac{1}{2}}{\frac{\sqrt{n'}}{2}}$. Отсюда следует:

$$s = \frac{n' - 1 - z_{кр} \cdot \sqrt{n'}}{2}. \qquad (3.9)$$

Эта приближенная оценка удобна при «ручной» обработке больших выборок.

Пример 2. Может ли медиана случайной величины, представленной выборкой (1.9), быть равна нулю? Уровень значимости α = 0,05.

Решение. Основная гипотеза H_0: Me = 0, альтернативная — H_1: Me ≠ 0. В выборке (1.9) имеется одно значение «0». Его отбрасываем. Остаются $n_+ = 18$ положительных и $n_- = 31$ отрицательных значений.

Поскольку значение $n' = n_+ + n_- = 49$ велико, критические точки можно оценить по (3.9). Двусторонней альтернативной гипотезе Me ≠ 0 на уровне значимости α = 0,05 соответствует $z_{кр} = 1{,}960$, значит: $s \approx \dfrac{49 - 1 - 1{,}96 \cdot \sqrt{49}}{2} \approx 18$.

В критическую область попадают по 18 крайних значений выборки. 18-е крайние значения «–4» и «1». Значит, доверительный интервал для медианы генеральной совокупности –4 < Me < 1 включает предполагаемое значение «0». Принимаем основную гипотезу H_0: Me = 0.

***Ответ*:** да, на уровне значимости α = 0,05 принимается H_0: Me = 0.

Замечание. На практике критерий знаков обычно применяется при сравнении двух генеральных совокупностей по попарно связанным выборкам. Например, чтобы сравнить работу двух приборов X и Y, на каждом из них измеряют одну и ту же серию образцов. В результате получают связанные пары значений X_i и Y_i. Приборы работают одинаково, когда разности $X_i - Y_i$ с одинаковой вероятностью принимают положительные и отрицательные значения (подсчет «плюсовых» и «минусовых» разностей дал название критерию знаков). Далее проверяется H_0: Me = 0 для выборки разностей $X_i - Y_i$.

4. ПРОВЕРКА ГИПОТЕЗ О ГЕНЕРАЛЬНЫХ СРЕДНИХ, ГЕНЕРАЛЬНЫХ ДИСПЕРСИЯХ И О СООТВЕТСТВИИ

4.1. Проверка гипотез о генеральных средних. t-Критерий Стьюдента: одновыборочный, двухвыборочный парный и непарный

В пункте 3.4 было показано, что для проверки статистических гипотез о математических ожиданиях случайных величин может использоваться Z-критерий, статистика которого родственна интервальной оценке (2.6): $\mu = \bar{x} \pm z_\gamma \cdot \frac{\sigma}{\sqrt{n}}$. Однако применять Z-критерий можно только в двух случаях. Во-первых, когда исследуемые генеральные совокупности распределены нормально и известны точные значения их дисперсий σ^2. И, во-вторых, когда имеющиеся выборки настолько велики, что их исправленные дисперсии s^2 практически сходятся к генеральным σ^2, а распределение выборочного среднего $\overline{X}$ становится нормальным, даже если распределение самой исследуемой случайной величины X несколько отличается от нормального.

На практике же часто встречаются ситуации, когда имеющиеся в распоряжении исследователя выборки малы, точные значения генеральной дисперсии σ^2 неизвестны, однако есть основания считать распределение генеральной совокупности нормальным (например, выполняются условия центральной предельной теоремы). В таких случаях справедлива интервальная оценка математического ожидания исследуемой случайной величины вида (2.7): $\mu = \bar{x} \pm t_{\gamma,\nu} \cdot \frac{s}{\sqrt{n}}$. В критерии для проверки гипотез о математических ожиданиях также следует заменить Z-статистику (3.2) на T-статистику вида (2.8). Поскольку T-статистика подчиняется распределению Стьюдента (2.9), то и критерий называется t-критерием Стьюдента. По своей сути t-критерий очень близко напоминает Z-критерий.

Одновыборочный критерий Стьюдента применяется, когда имеется выборка из n значений исследуемой случайной величины X и требуется проверить равенство математического ожидания $\mu(X)$ этой величины какому-либо наперед заданному значению m. Основная гипотеза H_0: $\mu = m$. Мерой различий между выборочным средним $\overline{X}$ и предполагаемым значением m служит T-статистика:

$$T = \frac{\overline{X} - m}{\frac{s}{\sqrt{n}}}. \qquad (4.1)$$

В отличие от Z-статистики (3.2) она использует оценку s, а не истинное значение σ среднеквадратичного отклонения случайной величины X.

А величина s тоже зависит от случайной выборки. По этой причине Т-статистика подчиняется распределению Стьюдента (2.9). Напомним, что распределение Стьюдента зависит от числа степеней свободы (2.10): $\nu = n - 1$, где n — объем используемой выборки. С ростом n распределение Стьюдента стремится к стандартному нормальному $N(t; 0; 1)$, как показано на рис. 2.3. Значит, и критерий Стьюдента при $n \to \infty$ должен давать такой же результат, как Z-критерий.

Если основная гипотеза H_0: $\mu = m$ верна, то значения Т-статистики, вычисленные по данным различных выборок одинакового объема n, должны группироваться возле $t = 0$ в соответствии с распределением Стьюдента (рис. 4.1). Чем больше абсолютное значение t, тем существеннее различие между выборочным средним $\bar{x}$ и предполагаемым математическим ожиданием m генеральной совокупности. В пользу двусторонней альтернативной гипотезы H_1: $\mu \neq m$ одинаково говорят большие различия $\bar{x} - m$ обоих знаков, поэтому критическая область состоит из двух симметричных «хвостов» распределения Стьюдента (рис. 4.1, *a*). Положение двусторонних критических точек $\pm t_{кр}$ таково, чтобы площадь каждого «хвоста» распределения была равна половине требуемого уровня значимости. Эта площадь равна вероятности ошибки 1-го рода — случайного превышения Т-статистикой критического значения $\pm t_{кр}$.

При односторонней альтернативной гипотезе H_1: $\mu < m$ или H_1: $\mu > m$ (рис. 4.1, *b*) критическая область состоит из одного «хвоста» распределения Стьюдента. Площадь такого «хвоста» равна уровню значимости. Поэтому между односторонними критическими точками $t_{кр1}$ и двусторонними $t_{кр}$ существует соотношение:

$$t_{кр1}(\alpha) = t_{кр}(2\alpha).$$

Это соотношение бывает полезно при пользовании таблицами критических точек распределения Стьюдента (прил. 4), в которых обычно приводятся двусторонние значения. Например, одностороннее критическое значение с $\alpha = 0{,}05$ равно затабулированному двустороннему с $\alpha = 0{,}1$.

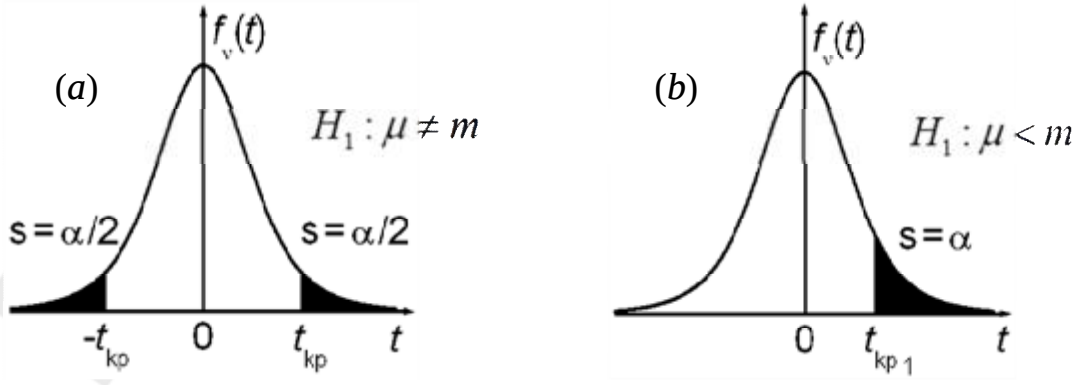

Рис. 4.1. График $f_\nu(t)$ распределения Стьюдента с $\nu = n - 1$ степенями свободы. Заштрихованы участки, соответствующие критической области двусторонней (*a*); правосторонней (*b*)

Пример 1. Генеральная совокупность X распределена нормально. Вариационный ряд выборки: –1, 0, 0, 0, 1, 1, 2, 3, 3, 7. Можно ли на уровне значимости $\alpha = 0{,}05$ считать, что (а) $\mu \neq 5$; (б) $\mu < 5$?

Решение. Основная гипотеза H_0: $\mu = 5$, альтернативные (а) H_1: $\mu \neq 5$ и (б) H_1: $\mu < 5$. Найдем характеристики выборки, необходимые для подсчета Т-статистики: объем $n = 10$. Среднее: $\bar{x} = \dfrac{-1+3\cdot 0+2\cdot 1+2+2\cdot 3+7}{10} = 1{,}6$.

Средний квадрат: $\overline{x^2} = \dfrac{(-1)^2+3\cdot 0^2+2\cdot 1^2+2^2+2\cdot 3^2+7^2}{10} = 7{,}4$. Исправленная дисперсия: $s^2 = \dfrac{10}{10-1}\cdot\left[\overline{x^2}-(\bar{x})^2\right] = 10\cdot\dfrac{7{,}4-1{,}6^2}{9} = \dfrac{48{,}4}{9}$. Значит, $s = \dfrac{\sqrt{48{,}4}}{3}$.

Т-статистика (4.1): $t = \left(\bar{x}-m\right)\cdot\dfrac{\sqrt{n}}{s} = (1{,}6-5)\cdot\dfrac{\sqrt{10}}{\sqrt{48{,}4}/3} = -\dfrac{3{,}4\cdot 3}{2{,}2} \approx -4{,}6$.

Критические точки распределения Стьюдента с $\nu = 10 - 1 = 9$ степенями свободы, соответствующие уровню значимости $\alpha = 0{,}05$: двусторонняя $t_{кр} = 2{,}26$, односторонняя $t_{кр1} = 1{,}83$.

(а) Значение Т-статистики $t \approx -4{,}6$ попадает в левый «хвост» двусторонней критической области: $t < -t_{кр}$, значит, принимаем H_1: $\mu \neq 5$.

(б) Альтернативная гипотеза H_1: $\mu < 5$ — левосторонняя, критическая область: $t < -t_{кр1} = -1{,}83$. Значение $t \approx -4{,}6$ попадает в нее, поэтому принимаем H_1: $\mu < 5$.

***Ответ*:** (а) да, $\mu \neq 5$; (б) да, $\mu < 5$.

Двухвыборочный парный критерий Стьюдента применяется, если две исследуемые выборки $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$ содержат связанные, парные значения, например значения какого-то показателя до (x_i) и после (y_i) воздействия. Составим выборку разностей парных значений: $d_1 = x_1 - y_1$, $d_2 = x_2 - y_2, \ldots, d_n = x_n - y_n$. Если случайные величины X и Y распределены нормально, то и их разность $X - Y$ также имеет нормальное распределение. Значит, используя t-критерий, по выборке разностей d_i можно проверить, значимо ли изменение показателя в результате воздействия.

Основную гипотезу H_0: $\mu(X) = \mu(Y)$ преобразуем к виду $\mu(X - Y) = 0$. Она проверяется против альтернативной двусторонней H_1: $\mu(X - Y) \neq 0$ или односторонних H_1: $\mu(X - Y) > 0$ или H_1: $\mu(X - Y) < 0$. Т-статистика (4.1) для разностей парных значений двух связанных выборок принимает вид:

$$T = \frac{\bar{d}}{\dfrac{s_d}{\sqrt{n}}}, \qquad (4.2)$$

где $\overline{d} = \overline{x} - \overline{y}$ — среднее значение, а $s_d = \sqrt{s_X^2 + s_Y^2}$ — исправленное среднеквадратичное отклонение выборки разностей $d_i = x_i - y_i$.

Найденное по выборке разностей значение Т-статистики (4.2) сравнивают с критическими точками распределения Стьюдента с $\nu = n - 1$ степенями свободы.

Пример 2. В рацион девяти лабораторных животных ввели БАД. Средний прирост массы тела составил 12 г при среднеквадратичном отклонении 5 г. Значимо ли воздействие БАД на уровне значимости 5 %?

Решение. Проверяем основную гипотезу H_0: $\mu = 0$ против двусторонней альтернативной H_1: $\mu \neq 0$. Считаем, что масса тела лабораторных животных подчиняется нормальному распределению. Это позволяет использовать парный двухвыборочный критерий Стьюдента. Вычислим его Т-статистику (4.2): $t = \dfrac{\overline{d} \cdot \sqrt{n}}{s_d} = \dfrac{12 \cdot \sqrt{9}}{5} = 7{,}2$.

Двусторонняя критическая точка распределения Стьюдента с $\nu = 9 - 1 = 8$ степенями свободы, соответствующая уровню значимости $\alpha = 0{,}05$, равна (прил. 4) $t_{кр} = 2{,}306$. Значение Т-статистики попадает в правый «хвост» критической области $t > t_{кр}$, значит, принимаем H_1: $\mu \neq 0$.

***Ответ*:** да, значимо.

Двухвыборочный непарный критерий Стьюдента применяется для сравнения математических ожиданий $\mu(X)$ и $\mu(Y)$ двух нормально распределенных величин X и Y с неизвестными одинаковыми дисперсиями $\sigma^2(X) = \sigma^2(Y) = \sigma^2$. Генеральные совокупности представлены не связанными между собой, непарными выборками X_i объемом n_X и Y_i объемом n_Y. Основная гипотеза H_0: $\mu(X) = \mu(Y)$ предполагает равенство генеральных средних $\mu(X)$ и $\mu(Y)$. Альтернативная гипотеза может быть двусторонней H_1: $\mu(X) \neq \mu(Y)$ или односторонней H_1: $\mu(X) > \mu(Y)$ (или H_1: $\mu(X) < \mu(Y)$).

Поскольку дисперсия σ^2 считается одинаковой для обеих генеральных совокупностей, стоит оценить ее более точно, используя данные обеих выборок одновременно. Число степеней свободы при вычислении исправленной дисперсии s_X^2 по выборке X_i равно $\nu_X = n_X - 1$, а сумма квадратов отклонений от среднего составляет $\sum_{i=1}^{n_X} (x_i - \overline{x})^2 = (n_X - 1) \cdot s_X^2$. И при вычислении исправленной дисперсии s_Y^2 по выборке Y_i число степеней свободы равно $\nu_Y = n_Y - 1$, а сумма квадратов отклонений от среднего $\sum_{i=1}^{n_Y} (y_i - \overline{y})^2 = (n_Y - 1) \cdot s_Y^2$. Поэтому общее число степеней свободы для объединенной оценки дисперсии:

$$\nu = n_X + n_Y - 2. \qquad (4.3)$$

Тогда несмещенная оценка дисперсии, общая для обеих выборок:

$$s^2 = \frac{\sum \left(x_i - \overline{x}\right)^2 + \sum \left(y_i - \overline{y}\right)^2}{(n_X - 1) + (n_Y - 1)},$$

или, используя исправленные дисперсии обеих выборок:

$$s^2 = \frac{(n_X - 1) \cdot s_X^2 + (n_Y - 1) \cdot s_Y^2}{n_X + n_Y - 2}. \quad (4.4)$$

Т-статистика должна быть мерой различия исследуемых μ(X) и μ(Y). Поэтому в ее числитель идет разница средних $\overline{X} - \overline{Y}$. А чтобы учесть разброс данных, в знаменателе Т-статистики должна быть оценка среднеквадратичного отклонения числителя $\sigma\left(\overline{X} - \overline{Y}\right)$. Исследуем дисперсию числителя, учитывая общее свойство дисперсии $D(X \pm Y) = D(X) + D(Y)$, определение дисперсии среднего (1.14) $D\left(\overline{X}\right) = \frac{D(X)}{n}$ и условие $\sigma^2(X) = \sigma^2(Y) = \sigma^2$. Тогда получим:

$$\sigma^2\left(\overline{X} - \overline{Y}\right) = \sigma^2\left(\overline{X}\right) + \sigma^2\left(\overline{Y}\right) = \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y} = \sigma^2 \cdot \left(\frac{1}{n_X} + \frac{1}{n_Y}\right).$$

Поскольку истинное значение σ^2 неизвестно, воспользуемся его оценкой (4.4). Тогда в случае двухвыборочного непарного критерия Стьюдента Т-статистика принимает вид:

$$T = \frac{\overline{X} - \overline{Y}}{s \cdot \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \quad (4.5)$$

где оценка s одинакового среднеквадратичного отклонения величин X и Y вычисляется в соответствии с формулой (4.4):

$$s = \sqrt{\frac{(n_X - 1) \cdot s_X^2 + (n_Y - 1) \cdot s_Y^2}{n_X + n_Y - 2}}.$$

Пример 3. Средние продажи препарата A за 12 дней составили 112 уп./день с исправленным среднеквадратичным отклонением s_A = 24 уп./день. Средние продажи препарата B за 6 дней — 78 уп./день при s_B = 32 уп./день. Можно ли на уровне значимости α = 0,05 утверждать, что препарат A продается лучше, чем препарат B?

Решение. Основная гипотеза H_0: μ(A) = μ(B), альтернативная H_1: μ(A) > > μ(B) — правосторонняя. Предположим, что дневные продажи обоих препаратов распределены нормально с одинаковой дисперсией — это позволит использовать двухвыборочный непарный критерий Стьюдента.

Объединенное число степеней свободы (4.3) равно: ν = 12 + 6 – 2 = 16. Правосторонняя критическая точка распределения Стьюдента при ν = 16 и α = 0,05 равна $t_{кр1}(16; 0{,}5) = t_{кр}(16; 0{,}1) = 1{,}75$.

Оценим общую дисперсию по (4.4):

$$s^2 = \frac{(12-1)\cdot 24^2 + (6-1)\cdot 32^2}{12+6-2} = 716 .$$

Т-статистика (4.5): $t = \dfrac{112-78}{\sqrt{716}\cdot\sqrt{\dfrac{1}{12}+\dfrac{1}{6}}} = \dfrac{34}{\sqrt{\dfrac{716\cdot(12+6)}{12\cdot 6}}} = \dfrac{34}{\sqrt{179}} \approx 2{,}54 .$

Полученное значение $t = 2{,}54$ попадает в правостороннюю критическую область $t > t_{кр1}$, поэтому принимаем альтернативную гипотезу H_1: μ(A) > μ(B).

***Ответ*:** да, препарат *A* продается лучше, чем препарат *B*.

4.2. Критерий Вилкоксона. Проверка гипотез о генеральных медианах

Критерий Вилкоксона — более мощный непараметрический критерий, сходный с критерием знаков (пункт 3.5). Увеличение мощности 1 – β имеет место благодаря использованию рангов выборочных значений. Обычно критерий Вилкоксона применяют как непараметрический аналог двухвыборочного парного критерия Стьюдента в тех случаях, когда исследуемые генеральные совокупности не подчиняются нормальному распределению. Этим обусловлено распространение термина «критерий Вилкоксона для связанных пар наблюдений».

Пусть имеются две выборки $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$, значения которых попарно связаны между собой. Это могут быть, например, значения какого-либо показателя, измеренные у одних и тех же людей до (x_i) и после (y_i) приема фармпрепарата. Составим выборку разностей: $d_1 = x_1 - y_1$, $d_2 = x_2 - y_2, \ldots, d_n = x_n - y_n$. Если препарат не имеет эффекта, то генеральные совокупности *X* и *Y* должны иметь одинаковое распределение. В этом случае должны быть соразмерны не только количества положительных n_+ и отрицательных n_- разностей, но и их абсолютные значения. На математическом языке это выражается нулевой гипотезой H_0: Me = 0, предполагающей равенство нулю медианы Me генеральной совокупности разностей *X* – *Y*. Альтернативная гипотеза — двусторонняя H_1: Me ≠ 0.

Как при пользовании критерием знаков, исключим из рассмотрения нулевые разности. Останется выборка разностей объема (3.6) $n' = n_+ + n_-$. Отранжируем ненулевые разности в порядке возрастания их модулей, т. е. минимальной разности припишем ранг 1, второй по абсолютному значению 2 и т. д., вплоть до максимальной разности, ранг которой равен n'. Затем сделаем ранги «знаковыми» — припишем каждому такой же знак, «+»

или «–», как у самой разности. Сумма знаковых рангов Rank_i служит статистикой критерия Вилкоксона:

$$W = \sum_{i=1}^{n'} \text{Rank}_i. \qquad (4.6)$$

Эта статистика дискретная, достигает крайних значений $\pm \frac{n' \cdot (n'+1)}{2}$, когда все ранги (и все разности) одного знака. Если справедлива основная гипотеза H_0, то распределение статистики (4.6) симметрично относительно нуля. Вероятности возможных значений W можно вычислить, полагая, что все комбинации знаковых рангов равновероятны. Всего имеется n' рангов. Каждый получает знак «+» или «–», поэтому всего возможно $2^{n'}$ комбинаций знаковых рангов. Затем определяют число комбинаций, которыми реализуется данное значение статистики W. Отношение этого числа к $2^{n'}$ равно вероятности данного значения. В критическую область включают крайние возможные значения W, пока сумма их вероятностей не достигнет требуемого уровня значимости α.

При большом объеме выборок (когда n' достигает нескольких десятков) распределение статистики (4.6) стремится к нормальному. В такой ситуации можно использовать критические точки Z-статистики, соответствующие требуемому уровню значимости α:

$$Z = \frac{|W| - \frac{1}{2}}{\sqrt{\frac{n' \cdot (n'+1) \cdot (2n'+1)}{6}}}. \qquad (4.7)$$

Пример. Можно ли считать воздействие значимым на уровне α = 0,05, если разница физиологического параметра, измеряемого у 10 лабораторных животных до и после этого воздействия, составила –40, –24, –22, –20, –13, –7, 0, 5, 16 и 76 единиц? Использовать критерий Вилкоксона.

Решение. Можно считать воздействие значимым, если удастся опровергнуть нулевую гипотезу H_0: Me = 0 и доказать H_1: Me ≠ 0.

Всего имеется n' = 9 ненулевых значений. Выстроим их в порядке возрастания абсолютного значения и присвоим знаковые ранги, как показано в таблице:

№	1	2	3	4	5	6	7	8	9
Значение	5	–10	–13	16	–20	–22	–24	–40	76
Знаковый ранг	1	–2	–3	4	–5	–6	–7	–8	9

Вычислим сумму знаковых рангов (4.6): W = 1 – 2 – 3 + 4 – 5 – 6 – 7 – – 8 + 9 = –17.

Определим критическую область, соответствующую уровню значимости α = 0,05. Всего возможно $2^{n'} = 2^9$ комбинаций знаковых рангов. Когда все ранги отрицательные, статистика (4.6) принимает свое минималь-

ное возможное значение $W = -1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 = -45$. Оно реализуется только при одной комбинации знаковых рангов, поэтому его вероятность равна $P(W = -45) = \frac{1}{2^9}$. Вероятность того, что статистика примет одно из значений $W = \pm 45$, в два раза больше: $P(W = \pm 45) = \frac{(1+1)}{2^9} = \frac{1}{256}$. Таким образом, двусторонняя критическая область $W = \pm 45$ отвечает уровню значимости $\alpha = \frac{1}{256} \approx 0{,}0039$.

Если минимальный ранг станет положительным, а все остальные останутся отрицательными, то статистика $W = +1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 = -43$. Значению $W = -43$ соответствует только одна комбинация рангов, значит, его вероятность $P(W = -43) = \frac{1}{2^9}$. Критическая область $|W| \geq 43$ состоит из четырех точек $W = \pm 45$ и $W = \pm 43$. Вероятность попадания в нее (уровень значимости) $P(|W| \geq 43) = \frac{1+1+1+1}{2^9} = \frac{2}{256} \approx 0{,}0078$.

Следующее возможное значение статистики $W = -41$ реализуется тоже только одной комбинацией рангов: $W = -1 + 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 = -41$. Поэтому вероятность попасть в одну из точек $W = \pm 41$ равна $P(W = \pm 41) = \frac{2}{2^9} = \frac{1}{256}$. А вероятность попасть в область $|W| \geq 41$ равна $P(|W| \geq 41) = \frac{2+1}{256} = \frac{3}{256}$.

Далее следует значение $W = -39$. Ему соответствуют две комбинации рангов: $W = -1 - 2 + 3 - 4 - 5 - 6 - 7 - 8 - 9 = -39$ и $W = +1 + 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 = -39$. Учитывая еще и точку $W = 39$, находим вероятность попасть в область $P(|W| \geq 39) = \frac{3+2}{256} = \frac{5}{256} \approx 0{,}0195$. Продолжая далее, находим, что значению $W = -37$ также соответствуют две комбинации: $W = -1 - 2 - 3 + 4 - 5 - \ldots$ и $W = +1 - 2 + 3 - 4 - \ldots = -37$. Поэтому $P(|W| \geq 37) = \frac{5+2}{256} = \frac{7}{256} \approx 0{,}0273$. Значение $W = -35$ реализуется тремя способами: $W = -1 - 2 - 3 - 4 + 5 - 6 - \ldots$, $W = +1 - 2 - 3 + 4 - 5 - \ldots$ и $W = -1 + 2 + 3 - 4 - \ldots$, значит $P(|W| \geq 35) = \frac{10}{256} \approx 0{,}0391$.

А следующее возможное значение статистики $W = -33$ может быть реализовано уже четырьмя комбинациями знаковых рангов: $W = \ldots - 5 + 6 - 7 - \ldots$, $W = +1 - 2 - 3 - 4 + 5 - 6 - \ldots$, $W = +1 + 2 - 3 + 4 - 5 - \ldots$ и

$W = +1 + 2 + 3 - 4 - \ldots$. Вероятность попасть в область $|W| \geq 33$ составляет $P(|W| \geq 33) = \frac{14}{256} \approx 0{,}0547$. Это значение уже превышает допустимый уровень значимости $\alpha = 0{,}05$. Поэтому следует заключить, что требуемому $\alpha = 0{,}05$ отвечает критическая область $|W| \geq 35$. Вычисленное выше значение статистики $W = -17$ в нее не попадает. Значит, опровержение основной гипотезы H_0: Me = 0 не найдено и ее следует принять.

***Ответ*:** нет, воздействие не значимо.

4.3. Проверка гипотез о генеральных дисперсиях. F-критерий Фишера

При решении различных задач бывает важно оценить не только положение случайных величин на числовой оси, но и соотношение их показателей рассеяния. Например, одно из условий применимости непарного двухвыборочного критерия Стьюдента (пункт 4.1) состоит в равенстве дисперсий исследуемых случайных величин. Проверка статистических гипотез о дисперсиях нормально распределенных генеральных совокупностей проводится с помощью *F-критерия Фишера*.

Сравним неизвестные дисперсии $\sigma^2(X)$ и $\sigma^2(Y)$ двух нормально распределенных случайных величин X и Y, представленных выборками объема n_X и n_Y. Основная гипотеза H_0: $\sigma^2(X) = \sigma^2(Y)$ заключается в равенстве этих неизвестных дисперсий. Судить об их значениях мы можем по величине исправленных выборочных дисперсий S_X^2 и S_Y^2. Эти оценки (S_X^2 и S_Y^2) сами зависят от случайной выборки, поэтому их значения будут несколько отличаться друг от друга, даже когда справедлива основная гипотеза H_0: $\sigma^2(X) = \sigma^2(Y)$. Мерой различий служит *F-статистика*, равная отношению большей исправленной выборочной дисперсии к меньшей. Пусть, для определенности, $S_X^2 > S_Y^2$, тогда:

$$F = \frac{S_X^2}{S_Y^2}. \qquad (4.8)$$

При справедливой основной гипотезе H_0 F-статистика подчиняется распределению Фишера–Снедекора. Это распределение зависит от двух параметров. Первый из них — число степеней свободы числителя F-статистики (большей исправленной выборочной дисперсии S_X^2):

$$\nu_X = n_X - 1. \qquad (4.9а)$$

А число степеней свободы знаменателя F-статистики (меньшей исправленной выборочной дисперсии S_Y^2) служит вторым параметром:

$$\nu_Y = n_Y - 1. \qquad (4.9б)$$

Плотность вероятности распределения Фишера–Снедекора (справочно):

$$f(F) = A \cdot \frac{F^{\frac{\nu_X - 2}{2}}}{(\nu_X \cdot F + \nu_Y)^{\frac{\nu_X + \nu_Y}{2}}},$$

где A — нормирующий множитель, ν_X — число степеней свободы числителя, а ν_Y — число степеней свободы знаменателя F-статистики. График распределения Фишера–Снедекора показан на рис. 4.2.

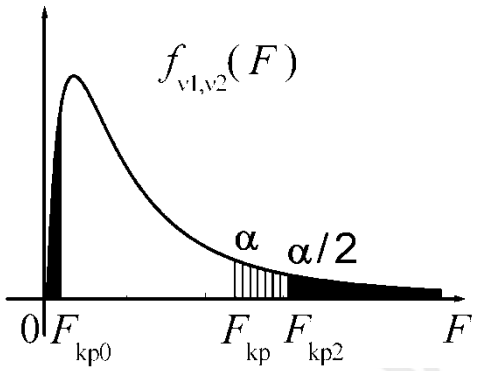

Рис. 4.2. График $f_{\nu_X;\nu_Y}(F)$ распределения Фишера–Снедекора с $\nu_X = n_X - 1$ и $\nu_Y = n_Y - 1$ степенями свободы. Заштрихованы участки, соответствующие односторонней и двусторонней критическим областям

Значение F-статистики всегда больше единицы (потому что в ее числитель всегда идет большая из двух исправленных выборочных дисперсий). Поэтому чаще выдвигается правосторонняя альтернативная гипотеза H_1: $\sigma^2(X) > \sigma^2(Y)$. А в таблицах обычно содержатся правосторонние критические точки $F_{кр}(\alpha, \nu_X; \nu_Y)$ распределения Фишера–Снедекора, отсекающие на графике распределения правостороннюю критическую область $F > F_{кр}$. Площадь этого «хвоста» равна требуемому уровню значимости α.

Если же альтернативная гипотеза двусторонняя H_1: $\sigma^2(X) \neq \sigma^2(Y)$, то α разделится пополам между двумя «хвостами» двусторонней критической области. Поэтому двустороннее критическое значение $F_{кр2}$ можно найти по таблице односторонних критических точек $F_{кр}$, отвечающих вдвое меньшему уровню значимости $\frac{\alpha}{2}$. Это соответствие можно записать так:

$$F_{кр2}(\alpha, \nu_X; \nu_Y) = F_{кр}(\frac{\alpha}{2}; \nu_X; \nu_Y).$$

Пример. Проверить на уровне значимости α = 0,05 правомерность предположения о равенстве дисперсий среднедневных продаж препаратов *A* и *B*, использовавшегося при решении примера 3 из пункта 4.1. (По условию примера 3 $s_A = 24$ и $s_B = 32$, количество дней наблюдения $n_A = 12$ и $n_B = 6$.)

Решение. Основная гипотеза H_0: $\sigma^2(A) = \sigma^2(B)$. Большая выборочная исправленная дисперсия у препарата B. Поэтому альтернативная гипотеза H_1: $\sigma^2(B) > \sigma^2(F)$.

Вычислим F-статистику (4.8): $F = \frac{s_B^2}{s_A^2} = \frac{32^2}{24^2} = \frac{16}{9} \approx 1{,}78$.

Число степеней свободы числителя: $\nu_B = 6 - 1 = 5$, число степеней свободы знаменателя: $\nu_A = 12 - 1 = 11$. В прил. 5 находим одностороннюю критическую точку $F_k(0{,}05; 5; 11) = 3{,}20$. Критическая область: $F > 3{,}20$.

Найденное значение статистики $F = 1{,}78$ не попадает в критическую область, поэтому считаем справедливой основную гипотезу H_0: $\sigma^2(A) = \sigma^2(B)$.

***Ответ*:** да, предположение $\sigma^2(A) = \sigma^2(B)$ верно при $\alpha = 0{,}05$.

4.4. Проверка гипотез об эквивалентности распределений. Критерий согласия Пирсона χ^2 (хи-квадрат)

Статистическая *эквивалентность* случайных величин подразумевает равенство (эквивалентность) их законов распределения. Пусть неизвестное распределение случайной величины X представлено выборкой X_1, X_2, …, X_n. Оно может быть признано статистически эквивалентным, во-первых, какому-либо распределению $F(x)$, позаимствованному из теории вероятностей и, во-вторых, другому неизвестному распределению случайной величины Y, представленной выборкой Y_1, Y_2, …, Y_n. Наблюдающиеся различия эмпирических распределений выборок статистически эквивалентных генеральных совокупностей должны быть случайными, несущественными. В силу закона больших чисел эти случайные различия должны становиться менее заметными при увеличении объема выборок. Статистические гипотезы об эквивалентности распределений проверяются при помощи *критериев согласия*.

Остановимся подробнее на проверке эквивалентности неизвестного распределения, представленного выборкой X_1, X_2, …, X_n, и теоретического распределения с известной интегральной функцией $F(x)$. Основная гипотеза H_0 предполагает равенство этих двух распределений.

Критерий Пирсона χ^2 (читается «хи-квадрат», от греческой буквы χ — «хи») основан на проверке согласия наблюдаемых частот m_i теоретическим вероятностям p_i. Если вероятность i-го значения дискретной случайной величины равна p_i, то в сериях из n независимых испытаний это значение должно появляться в среднем по $n \cdot p_i$ раз ($n \cdot p_i$ — математическое ожидание биномиального распределения). Значит, «теоретическая» частота появления i-го значения в выборке объема n равна $n \cdot p_i$. Мерой отклонения наблюдаемой частоты m_i от теоретического предсказания

$n \cdot p_i$ в критерии Пирсона служит величина $\frac{(m_i - n \cdot p_i)^2}{n \cdot p_i}$. Просуммировав ее по всем k возможным значениям, получим статистику χ^2 для дискретной случайной величины:

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - n \cdot p_i)^2}{n \cdot p_i}. \qquad (4.10а)$$

Если частоты m_i вариант выборки малы (например, из-за того, что исследуемая случайная величина непрерывна), необходимо сгруппировать значения выборки x_i по l частичным интервалам $[h_0; h_1)$, $[h_1; h_2)$, …, $[h_{l-1}; h_l)$, как при построении гистограммы (пункт 1.4). В результате группировки получим суммарные частоты интервалов m_1^*, m_2^*, …, m_n^*. «Теоретические» вероятности p_i попадания в эти интервалы легко найти по интегральной функции $F(x)$ предполагаемого «теоретического» распределения, используя соотношение $p_i = F(h_i) - F(h_{i-1})$.

Непрерывные «теоретические» распределения $F(x)$ обычно определены на всей числовой оси (при этом $F(-\infty) = 0$ и $F(+\infty) = 1$). Поэтому крайние интервалы также должны простираться до $\pm\infty$. Тогда «теоретическая» вероятность попадания в первый интервал $-\infty < x < h_1$ равна:

$$p_1 = F(h_1) - F(-\infty) = F(h_1).$$

Соответственно, для последнего интервала $h_{l-1} \le x < +\infty$ получаем вероятность: $p_l = F(+\infty) - F(h_{l-1}) = 1 - F(h_{l-1})$.

А статистика χ^2 для выборки, сгруппированной по l частичным интервалам, примет вид:

$$\chi^2 = \sum_{i=1}^{l} \frac{(m_i^* - n \cdot p_i)^2}{n \cdot p_i}. \qquad (4.10б)$$

Когда справедлива нулевая гипотеза H_0, неизвестное распределение исследуемой случайной величины эквивалентно предполагаемому $F(x)$. В этой ситуации наблюдаемые частоты m_i^* не должны сильно отличаться от «теоретических» $n \cdot p_i$, а вычисленное по (4.10) значение статистики χ^2 будет небольшим. Если же основная гипотеза H_0 неверна, сумма квадратов отклонений в (4.10) достигнет больших, но обязательно положительных значений. Поэтому критерий Пирсона χ^2 используется с правосторонней альтернативной гипотезой H_1. Она принимается, когда $\chi^2 \ge \chi^2_{кр}$. (Таблицу правосторонних критических точек $\chi^2_{кр}$ можно найти в прил. 6.)

Распределение статистики χ^2 зависит от числа степеней свободы:

$$\nu = l - 1 - r, \qquad (4.11)$$

где l — количество интервалов, по которым группируются значения выборки. (Если группировка не проводится, то вместо l используют число возможных значений k исследуемой дискретной генеральной совокупно-

сти). Единица вычитается, поскольку на значения наблюдаемых частот налагается одно ограничение: $\sum m_i^* = n$. Также вычитается r — количество параметров «теоретического» распределения, значения которых оценивают по выборке. Пусть, например, требуется проверить соответствие исследуемой генеральной совокупности нормальному закону. Если в качестве параметров нормального распределения μ и σ используют выборочное среднее $\bar{x}$ и исправленное среднеквадратичное отклонение s, то в (4.11) следует подставлять $r = 2$. Если у нормального закона зафиксировано значение одного параметра, например μ, то надо использовать $r = 1$. (Такая ситуация может возникнуть, например, когда исследуемая случайная величина с равной вероятностью принимает положительные и отрицательные значения. Поэтому зафиксировано μ = 0. А в формулу нормального распределения подставляют только один параметр выборки s.)

Если нулевая гипотеза H_0 справедлива, то при увеличении объема выборки ($n \to \infty$) распределение статистики (4.10) приближается к так называемому распределению Пирсона. Критические точки именно этого предельного распределения представлены в статистических таблицах. Его плотность вероятности (справочно):

$$f(u) = A \cdot e^{-\frac{u}{2}} \cdot u^{\frac{\nu}{2}-1},$$

где A — нормировочный множитель, ν — число степеней свободы. Типичный график распределения Пирсона χ^2 показан на рис. 4.3.

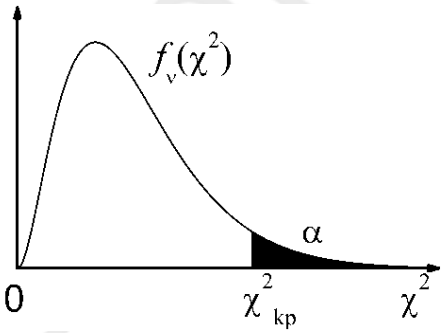

Рис. 4.3. График $f_\nu(\chi^2)$ распределения Пирсона $\chi^2_{кр}$ с $\nu = l - 1 - r$ степенями свободы. Заштрихован участок, соответствующий правосторонней критической области

«Предельные» критические значения $\chi^2_{кр}$ не точны в случае малой выборки. Практическая применимость критерия Пирсона χ^2 зависит как от объема выборки n, так и от «теоретических» частот $n \cdot p_i$. Считается, что этот критерий можно использовать, когда, во-первых, $n \geq 50$ и, во-вторых, $n \cdot p_i \geq 5$. Даже если выборка достаточно велика, у некоторых интервалов группирования «теоретические» частоты $n \cdot p_i$ могут оказаться малыми. Такие интервалы следует объединять с их соседями. Заметим, что при

этом не только увеличатся интервальные частоты m_i^* и $n \cdot p_i$, но и уменьшится число степеней свободы (4.11).

Пример. Можно ли считать нормальным распределение случайной величины X, представленной выборкой (1.9)? Использовать значения выборочного среднего $\bar{x} = -2{,}72$ и исправленного среднеквадратичного отклонения $s = 9{,}36$. Требуемый уровень значимости $\alpha = 0{,}05$.

Решение. Основная гипотеза H_0 предполагает, что случайная величина X подчиняется нормальному распределению $N(x; -2{,}72; 9{,}36)$ с математическим ожиданием $\mu = -2{,}72$ и среднеквадратичным отклонением $\sigma = 9{,}36$. Объем выборки (1.9) $n = 50$ достаточен для применения критерия Пирсона χ^2. Этот критерий используется с правосторонней альтернативной гипотезой H_1.

В пункте 1.4 значения выборки (1.9) уже были сгруппированы по семи частичным интервалам. Получившийся интервальный ряд показан в табл. 1.5. Дополним его «теоретическими» частотами $n \cdot p_i$. Если нулевая гипотеза H_0 верна, то вероятность p_i равна разности значений функции Лапласа на границах i-го интервала:

$$p_i = \Phi\left(\frac{h_i - \mu}{\sigma}\right) - \Phi\left(\frac{h_{i-1} - \mu}{\sigma}\right) = \Phi\left(\frac{h_i + 2{,}72}{9{,}36}\right) - \Phi\left(\frac{h_{i-1} + 2{,}72}{9{,}36}\right).$$

Тогда для первого интервала:

$$p_1 = \Phi\left(\frac{-22 + 2{,}72}{9{,}36}\right) - \Phi(-\infty) \approx \Phi(-2{,}06) + \frac{1}{2} = \frac{1}{2} - \Phi(2{,}06) \approx$$

$$\approx 0{,}5 - 0{,}4803 = 0{,}0197.$$

Его «теоретическая» частота $n \cdot p_1 = 50 \cdot 0{,}0197 = 0{,}985$ мала.

Продолжим вычисления для остальных интервалов:

$$p_2 = \Phi\left(\frac{-16 + 2{,}72}{9{,}36}\right) - \Phi\left(\frac{-22 + 2{,}72}{9{,}36}\right) \approx \Phi(2{,}06) - \Phi(1{,}42) \approx$$

$$\approx 0{,}4803 - 0{,}4222 = 0{,}0581.$$

2-я «теоретическая» частота $n \cdot p_2 = 50 \cdot 0{,}0581 = 2{,}905$ мала.

$$p_3 = \Phi\left(\frac{-10 + 2{,}72}{9{,}36}\right) - \Phi(-1{,}42) \approx \Phi(1{,}42) - \Phi(0{,}78) \approx$$

$$\approx 0{,}4222 - 0{,}2823 = 0{,}1399.$$

3-я «теоретическая» частота $n \cdot p_3 = 50 \cdot 0{,}1399 = 6{,}995$.

$$p_4 = \Phi\left(\frac{-4 + 2{,}72}{9{,}36}\right) - \Phi(-0{,}78) \approx \Phi(0{,}78) - \Phi(0{,}14) \approx$$

$$\approx 0{,}2823 - 0{,}0557 = 0{,}2266.$$

4-я «теоретическая» частота $n \cdot p_4 = 50 \cdot 0{,}2266 = 11{,}33$.

$$p_5 = \Phi\left(\frac{2 + 2{,}72}{9{,}36}\right) - \Phi(-0{,}14) \approx \Phi(0{,}14) + \Phi(0{,}50) \approx 0{,}0557 + 0{,}1915 = 0{,}2472.$$

5-я «теоретическая» частота $n \cdot p_5 = 50 \cdot 0{,}2472 = 12{,}36$.

$$p_6 = \Phi\left(\frac{8+2{,}72}{9{,}36}\right) - \Phi(0{,}50) \approx \Phi(1{,}15) - \Phi(0{,}50) \approx$$
$$\approx 0{,}3749 - 0{,}1915 = 0{,}1834.$$

6-я «теоретическая» частота $n \cdot p_6 = 50 \cdot 0{,}1834 = 9{,}17$.

$$p_7 = \Phi(+\infty) - \Phi\left(\frac{8+2{,}72}{9{,}36}\right) \approx 0{,}5 - \Phi(1{,}15) \approx 0{,}5 - 0{,}3749 = 0{,}1251.$$

7-я «теоретическая» частота $n \cdot p_7 = 50 \cdot 0{,}1251 = 6{,}255$.

Сведем результаты в таблицу:

Интервал	$x < -22$	$[-22; -16)$	$[-16; -10)$	$[-10; -4)$	$[-4; 2)$	$[2; 8)$	$x \geq 8$
m_i^*	2	2	4	10	16	9	7
$n \cdot p_i$	0,985	2,905	6,995	11,33	12,36	9,17	6,255

Частоты первых двух интервалов малы, поэтому объединяем эти интервалы с третьим. Результат перегруппировки:

Интервал	$x < -10$	$[-10; -4)$	$[-4; 2)$	$[2; 8)$	$x \geq 8$
m_i^*	8	10	16	9	7
$n \cdot p_i$	10,885	11,33	12,36	9,17	6,255

Используя эти данные, вычислим значение статистики χ^2:

$$\chi^2 = \frac{(8-10{,}885)^2}{10{,}885} + \frac{(10-11{,}33)^2}{11{,}33} + \frac{(16-12{,}36)^2}{12{,}36} + \frac{(9-9{,}17)^2}{9{,}17} +$$
$$+ \frac{(7-6{,}225)^2}{6{,}225}.$$

$$\chi^2 \approx 0{,}765 + 0{,}156 + 1{,}072 + 0{,}003 + 0{,}096 = 2{,}092.$$

Новое число интервалов после перегруппировки $l = 5$. «Теоретические» частоты $n \cdot p_i$ получены с использованием двух параметров выборки, $\bar{x} = -2{,}72$ и $s = 9{,}36$, поэтому $r = 2$. Значит, число степеней свободы (4.11) $\nu = 5 - 1 - 2 = 2$. Правостороннюю критическую точку распределения χ^2 с двумя степенями свободы при $\alpha = 0{,}05$ находим в прил. 6: $\chi^2_{\text{кр}}(2; 0{,}05) \approx 5{,}991$.

Значение статистики $\chi^2 \approx 2{,}092$ меньше критического $\chi^2_{\text{кр}}(2; 0{,}05) \approx 5{,}991$, поэтому (рис. 4.3) принимаем нулевую гипотезу H_0.

***Ответ*:** да, распределение можно считать нормальным при $\alpha = 0{,}05$.

4.5. Критерий Колмогорова–Смирнова

Критерий Колмогорова–Смирнова используется для проверки статистических гипотез об эквивалентности распределений непрерывных случайных величин по эмпирическим функциям этих распределений (1.5).

Пусть известна эмпирическая функция $F^*(x)$, построенная по выборке значений непрерывной случайной величины X. А интегральная функция $F(x)$ описывает некоторое «теоретическое» распределение. В качестве примера на рис. 4.4 построены эмпирическая функция распределения выборки (1.9) и интегральная функция нормального распределения с параметрами $\mu = -2{,}72$ и $\sigma = 9{,}36$, как у выборки (1.9): $\bar{x} = -2{,}72$ и $s = 9{,}36$. Можно ли считать, что эмпирическая $F^*(x)$ согласуется с «теоретической» $F(x)$?

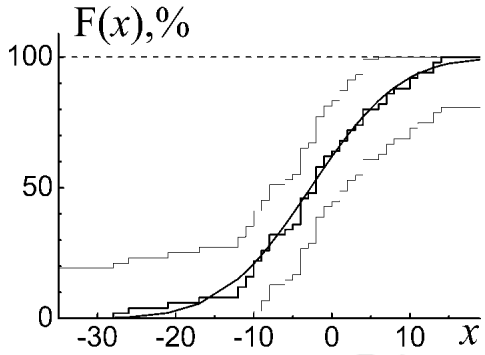

Рис. 4.4. Гладкая линия — интегральная функция $F_{Norm}(x)$ нормального распределения с $\mu = -2{,}72$ и $\sigma = 9{,}36$. Ломаными линиями показаны эмпирическая функция распределения $F^*(x)$ выборки (1.9) и полоса $F^*(x) \pm D_{кр}$, где $D_{кр}$ — критическое значение (4.13, *а*) статистики Колмогорова–Смирнова (4.12) для объема выборки $n = 50$ и уровня значимости $\alpha = 0{,}05$

Нулевая гипотеза H_0 предполагает, что распределение генеральной совокупности X полностью соответствует «теоретической» функции $F(x)$. Статистика критерия Колмогорова–Смирнова D равна максимальному модулю разности значений функций $F^*(x)$ и $F(x)$:

$$D = \max\left|F^*(x) - F(x)\right|. \qquad (4.12)$$

Когда эмпирическая $F^*(x)$ и «теоретическая» $F(x)$ функции распределения сильно отличаются, статистика (4.12) принимает большое положительное значение. Поэтому альтернативная гипотеза H_1 должна быть правосторонней. Ей соответствует правосторонняя критическая область: $D \geq D_{кр}$.

При увеличении объема выборки $n \to \infty$ распределение величины $D \cdot \sqrt{n}$ стремится к распределению Колмогорова–Смирнова. Вероятность того, что величина с таким распределением примет значение $D \cdot \sqrt{n} > K$, выражается бесконечным рядом (справочно):

$$P\left(D \cdot \sqrt{n} \geq K\right) = 2 \cdot \sum_{i=1}^{\infty} (-1)^{i-1} \cdot e^{-2i^2K^2} = 2 \cdot e^{-2K^2} - 2 \cdot e^{-8K^2} + 2 \cdot e^{-18K^2} - \ldots$$

В правосторонней критической точке $K = D_{кр} \cdot \sqrt{n}$ эта вероятность равна уровню значимости α. Поэтому критическое значение статистики (4.12) $D_{кр}$ и уровень значимости α связаны соотношением:

$$\alpha = 2 \cdot e^{-2nD_{кр}^2} - 2 \cdot e^{-8nD_{кр}^2} + 2 \cdot e^{-18nD_{кр}^2} - \ldots$$

Когда объем выборки n велик, абсолютные значения второго и последующих членов этого ряда малы по сравнению с первым слагаемым $2 \cdot e^{-2K^2}$. В этом случае ($n \to \infty$) можно оставить только первое слагаемое:

$$\alpha \approx 2 \cdot e^{-2nD_{кр}^2}.$$

Прологарифмировав левую и правую части этого равенства, можно найти зависимость правостороннего критического значения статистики Колмогорова–Смирнова (4.12) от объема выборки n и требуемого уровня значимости α:

$$D_{кр} = \sqrt{\frac{1}{2n} \cdot \ln\frac{2}{\alpha}}. \qquad (4.13а)$$

Вернемся к выборке (1.9). Ее объем n = 50. Если уровень значимости α = 0,05, то критическим будет значение статистики:

$$D_{кр} = \sqrt{\frac{1}{2 \cdot 50} \cdot \ln\frac{2}{0{,}05}} \approx 0{,}192.$$

Если значения сравниваемых функций распределения $F(x)$ и $F^*(x)$ где-либо отличаются больше, чем на $D_{кр}$, то основную гипотезу H_0 о соответствии исследуемой случайной величины X «теоретической» функции распределения $F(x)$ следует отвергнуть.

На рис. 4.4 дополнительно проведены две линии, отстоящие от графика эмпирической функции распределения $F^*(x)$ по вертикали на $\pm D_{кр}$ = ±19,2 %. Полоса, заключенная между ними, с доверительной вероятностью 1 – α накрывает истинную функцию распределения исследуемой генеральной совокупности. График «теоретической» функции $F(x)$ везде укладывается в эту полосу, значит, оснований отвергнуть нулевую гипотезу нет. Таким образом, на уровне значимости α = 0,05 выборка (1.9) соответствует нормальному распределению с математическим ожиданием μ = –2,72 и среднеквадратичным отклонением σ = 9,36. (В данном случае критерий Колмогорова–Смирнова привел нас к такому же заключению, что и критерий Пирсона χ^2.)

Критерий Колмогорова–Смирнова может использоваться и для проверки эквивалентности двух распределений, представленных выборками объема n_1 и n_2. В этом случае сравниваются эмпирические функции: $D = \max\left|F_1^*(x) - F_2^*(x)\right|$. А критическое значение статистики зависит от объемов обеих выборок:

$$D_{кр} = \sqrt{\frac{n_1 + n_2}{2n_1n_2} \cdot \ln\frac{2}{\alpha}}. \qquad (4.13б)$$

Замечание. Мощность критерия Колмогорова–Смирнова падает, если выборки малы. Даже при n = 50 и α = 0,05 для опровержения нулевой гипотезы требуется, чтобы функции распределений отличались хотя бы на $D_{кр}$ = 19,2 %. Но значения этих функций сами заключены в интервале от 0 до 100 %.

4.6. Сравнение нескольких групп

Все рассмотренные выше статистические гипотезы заключались в одиночных сравнениях. Одна генеральная совокупность (или ее параметр) сравнивалась с другой генеральной совокупностью (или ее параметром). Для проверки таких гипотез использовались двухвыборочные разновидности статистических критериев. Одновыборочные критерии применялись, когда место второй генеральной совокупности занимало «теоретическое» распределение (или предполагаемое значение параметра «теоретического» распределения). Отвергая или принимая основную статистическую гипотезу H_0, мы могли совершить ошибку первого или второго рода (пункт 3.1). Требовалось, чтобы вероятность ошибки 1-го рода (посчитать случайные различия значимыми) была меньше уровня значимости α.

Существуют ситуации, когда надо проводить сразу несколько сравнений. Например, чтобы найти наиболее эффективный фармакологический препарат среди нескольких имеющихся. Для этого можно попытаться попарно сравнить средние значения физиологического эффекта, вызываемого этими препаратами. Но каждое дополнительное сравнение увеличивает вероятность ошибки. Поэтому в каждом одиночном сравнении надо использовать меньший уровень значимости α_1, чтобы общий вывод соответствовал требуемому уровню значимости α.

Пусть всего требуется провести k одиночных сравнений. Вероятность допустить ошибку 1-го рода в одном таком сравнении должна быть меньше, чем α_1. Значит, минимальная вероятность не ошибиться при одиночном сравнении равна $1 - \alpha_1$, тогда вероятность не ошибиться ни разу в серии из k одиночных сравнений равна $(1 - \alpha_1)^k$. А вероятность ошибочности общего вывода равна $1 - (1 - \alpha_1)^k$. Именно эта вероятность и не должна превышать общего уровня значимости α. Поэтому α_1 и α можно связать соотношением:

$$\alpha = 1 - (1 - \alpha_1)^k. \qquad (4.14)$$

Выразим уровень значимости, требуемый для каждого одиночного сравнения:

$$\alpha_1 = 1 - \sqrt[k]{1 - \alpha}. \qquad (4.15а)$$

Точную формулу (4.15а) можно заменить более удобной для «ручных» вычислений. Раскроем скобки в формуле (4.14): $\alpha = 1 - 1 + k \cdot \alpha_1 - - \ldots - \alpha_1^k$. Поскольку $\alpha_1 << 1$, абсолютные значения всех слагаемых со степенями $\alpha_1^2, \ldots, \alpha_1^k$ много меньше, чем $k \cdot \alpha_1$. Отбросив эти слагаемые, получим: $\alpha \approx k \cdot \alpha_1$. Отсюда следует, что при k сравнениях уровень значимости для каждого из них должен быть в k раз меньше, чем требуется для общего вывода (поправка Бонферрони):

$$\alpha_1 = \frac{\alpha}{k}. \qquad (4.15б)$$

Например, если вывод надо сделать на основе пяти сравнений и с уровнем значимости α, то каждое одиночное сравнение следует проводить на уровне значимости $\frac{\alpha}{5}$. А это существенно снизит мощность используемых статистических критериев, т. е. повысит вероятность ошибки 2-го рода (существенные различия посчитать случайными). Для множественных сравнений разработаны специальные, более эффективные статистические критерии и методы (см. раздел 5).

С-критерий Кочрена* используется для сравнения дисперсий нескольких нормально распределенных случайных величин, представленных выборками одинакового объема n. Основная гипотеза H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ заключается в равенстве дисперсий всех k исследуемых генеральных совокупностей. Альтернативная гипотеза H_1 предполагает, что хотя бы одна генеральная дисперсия значительно больше других. Статистика критерия равна отношению наибольшей исправленной выборочной дисперсии $\max S^2$ к их общей сумме:

$$C = \frac{\max S^2}{S_1^2 + S_2^2 + \ldots + S_k^2}. \qquad (4.16)$$

Критическая область правосторонняя: $C \geq C_{\text{кр}}$. Критические точки $C_{\text{кр}}(\alpha; n; k)$ даны в прил. 7. Они зависят от требуемого уровня значимости α, объема выборок n и числа сравниваемых дисперсий k. Существует связь с критическими точками распределения Фишера–Снедекора $F_{\text{кр}}$:

$$C_{\text{кр}} = \frac{1}{1 + \frac{k-1}{F_{\text{кр}}}}. \qquad (4.17)$$

Здесь $F_{\text{кр}}\left(\frac{\alpha}{k}; \nu_X; \nu_Y\right)$ отвечают уровню значимости $\frac{\alpha}{k}$, число степеней свободы числителя равно $\nu_X = n - 1$ и число степеней свободы знаменателя равно $\nu_Y = (k-1)\cdot(n-1)$.

Пример. Исправленные дисперсии пяти выборок одинакового объема $n = 6$ равны: $S_1^2 = 30$, $S_2^2 = 5$, $S_3^2 = 4$, $S_4^2 = 9$, $S_5^2 = 7$. Проверить на уровне значимости $\alpha = 0{,}05$, могут ли эти выборки представлять нормальные распределения с одинаковыми дисперсиями.

Решение. Основная гипотеза H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$. Проверим ее с помощью критерия Кочрена. Правостороннее критическое значение распределения Кочрена, отвечающее уровню значимости $\alpha = 0{,}05$, объему

* (англ. Cochrane's C test) Этот критерий удобен, если статистический анализ проводится «вручную». Компьютерные статистические программы обычно используют для сравнения нескольких дисперсий другие, более оптимальные критерии (Levene's test, Bartlett's test).

выборки $n = 6$ и числу сравниваемых дисперсий $k = 5$, находим в прил. 7: $C_{кр}(0{,}05; 6; 5) \approx 0{,}506$.

Вычислим статистику (4.16). Максимальная выборочная исправленная дисперсия $\max S^2 = S_1^{\,2} = 30$, значит: $C = \dfrac{30}{30+5+4+9+7} = \dfrac{30}{55} \approx 0{,}545$.

Нулевую гипотезу отвергаем, так как С-статистика попала в критическую область $C \geq C_{кр}$.

***Ответ*:** нет, дисперсии не одинаковы.

5. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ используется для сравнения математических ожиданий (генеральных средних μ) сразу нескольких случайных величин. Суть метода состоит в анализе дисперсии имеющихся данных. Различают общую, факторную и остаточную дисперсии. Вывод о возможном равенстве математических ожиданий всех генеральных совокупностей делается на основе единственного сравнения факторной и остаточной дисперсий. Поэтому дисперсионный анализ не требует поправки Бонферрони (4.15). Это делает его гораздо более мощным статистическим критерием, чем t-критерий Стьюдента с поправкой Бонферрони.

5.1. Факторы, уровни факторов, группы

В основе метода дисперсионного анализа лежит модель случайной величины X, изменяющейся под действием каких-либо факторов. Эта случайная величина подчиняется нормальному распределению. Под воздействием факторов может изменяться только ее математическое ожидание $\mu(X)$, а дисперсия σ^2 всегда остается неизменной.

Терминология дисперсионного анализа имеет свои особенности. Любой параметр, который может повлиять на исследуемую случайную величину X, называется *фактором*. Например, содержание примеси (величина X может зависеть от таких факторов, как технология приготовления препарата, используемые прекурсоры и др.). *Уровни* фактора различаются интенсивностью или видом воздействия (метод очистки препарата от примесей, поставщик прекурсоров и т. д.).

Чтобы проверить, влияет ли интересующий нас фактор на случайную величину X, следует измерить ее значения при всевозможных уровнях этого фактора. Выборка, полученная при одном уровне фактора, называется *группой*. Предположим, что фактор A имеет k уровней: $A_1, A_2, \ldots, A_k$. Тогда каждому уровню A_j должна соответствовать своя группа значений $x_{1j}, x_{2j}, \ldots x_{ij}, \ldots$ Элементы этой группы обозначены двумя индексами. Первый индекс i равен номеру данного элемента внутри группы. Второй

индекс j указывает номер группы (уровень фактора). Для каждой группы можно найти групповое среднее:

$$\bar{x}_j = \frac{x_{1j} + x_{2j} + ... + x_{n_j j}}{n_j}. \quad (5.1)$$

и групповую исправленную дисперсию:

$$s_j^2 = \frac{(x_{1j} - \bar{x}_j)^2 + ... + (x_{n_j j} - \bar{x}_j)^2}{n_j - 1}. \quad (5.2)$$

Если фактор A не влияет на случайную величину X, то распределение этой случайной величины остается неизменным при любом уровне фактора. Группы будут отличаться друг от друга незначимо, поскольку все их элементы x_{ij} выбраны случайным образом из этого неизменного распределения. На рис. 5.1 эта ситуация схематично представлена левой колонкой (a, b, c). При всех трех изображенных уровнях фактора A_1 (рис. 5.1, *a*), A_2 (рис. 5.1, *b*) и A_3 (рис. 5.1, *c*) распределение случайной величины X одинаковое, как показано пунктирными линиями. Соответствующие группы обозначены гистограммами. Небольшие различия этих гистограмм отражают случайные, несущественные различия групп. Групповые средние $\bar{x}_1$, $\bar{x}_2$ и $\bar{x}_3$ несколько отличаются друг от друга (и от неизменного μ), однако эти различия не должны быть значимыми, большими.

Если же фактор A влияет на математическое ожидание μ(X), то при изменении уровня фактора распределение случайной величины X сдвигается вдоль числовой оси, что иллюстрирует правая колонка (d, e, f) на рис. 5.1. В этой ситуации различия между групповыми средними $\bar{x}_1$, $\bar{x}_2$ и $\bar{x}_3$ становятся значимыми, большими.

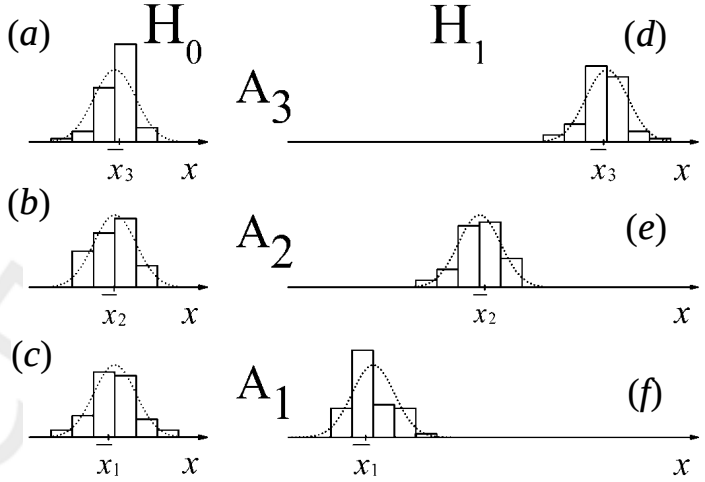

Рис. 5.1. Плотность вероятности нормально распределенной случайной величины X (пунктирная линия) и гистограммы групп ее значений, полученных при трех разных уровнях фактора A_1 (*c*, *f*), A_2 (*b*, *e*) и A_3 (*a*, *d*). В левой колонке (a, b, c) фактор себя не проявляет. В правой колонке (d, e, f) влияние фактора приводит к значимому изменению группового среднего $\bar{x}_j$

5.2. Дисперсия общая, факторная, остаточная

Дисперсионный анализ предназначен для проверки основной гипотезы H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, заключающейся в равенстве математических ожиданий μ_j случайной величины X при всех уровнях фактора A_j. Если эта гипотеза справедлива, то неизвестную дисперсию σ^2 случайной величины X можно оценить тремя различными способами.

Во-первых, можно объединить все группы в одну общую выборку. Ее объем N равен сумме объемов всех групп:

$$N = n_1 + n_2 + \ldots + n_k. \qquad (5.3)$$

Общее среднее $\bar{x}$ этой выборки найдем, используя групповые средние (5.1):

$$\bar{x} = \frac{\bar{x}_1 \cdot n_1 + \bar{x}_2 \cdot n_2 + \ldots + \bar{x}_k \cdot n_k}{N}. \qquad (5.4а)$$

Тогда *общая* исправленная дисперсия, вычисленная сразу по всем значениям x_{ij} объединенной выборки в соответствии с определением (1.19а), равна:

$$s^2_{\text{общ}} = \frac{1}{N-1} \cdot \sum_i \sum_j (x_{ij} - \bar{x})^2. \qquad (5.4б)$$

Число степеней свободы при вычислении общей дисперсии:

$$\nu_{\text{общ}} = \text{N} - 1. \qquad (5.5)$$

Во-вторых, можно использовать разброс групповых средних $\bar{x}_j$. Когда справедлива H_0 (фактор не влияет на случайную величину X), все группы являются выборками из одного и того же распределения. Разброс групповых средних мал. Это можно увидеть в левом столбце (a, b, c) на рис. 5.1. Предположим, для простоты, что объемы групп одинаковы: $n_1 + n_2 + \ldots + n_k = n$. Тогда между дисперсией группового среднего $\sigma^2(\bar{x})$ и дисперсией σ^2 случайной величины X выполняется соотношение (1.14): $\sigma^2(\bar{x}) = \frac{\sigma^2}{n}$. Из него получаем: $\sigma^2 = n \cdot \sigma^2(\bar{x})$.

Исправленная дисперсия (1.19а) среднего значения для групп равного объема n равна: $s^2(\bar{x}) = \frac{(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2 + \ldots + (\bar{x}_k - \bar{x})^2}{k-1}$. Домножив на n, получим оценку дисперсии σ^2 случайной величины X:

$$s^2_{\text{факт}} = \frac{n \cdot (\bar{x}_1 - \bar{x})^2 + \ldots + n \cdot (\bar{x}_k - \bar{x})^2}{k-1}.$$

Это выражение легко обобщить на случай неодинаковых объемов групп. Последовательно заменив n на n_1, n_2, …, n_k, дадим определение *факторной* исправленной дисперсии:

$$s^2_{\text{факт}} = \frac{n_1 \cdot (\bar{x}_1 - \bar{x})^2 + n_2 \cdot (\bar{x}_2 - \bar{x})^2 + ... + n_k \cdot (\bar{x}_k - \bar{x})^2}{k-1}. \qquad (5.6)$$

Эта оценка дисперсии σ^2 называется факторной, поскольку она выявляет влияние фактора на случайную величину X. Если это влияние существенно, то разброс групповых средних $\bar{x}_j$ будет большим, как показано в правой колонке (d, e, f) на рис. 5.1. Соответственно вырастут слагаемые в числителе (5.6) и сама $s^2_{\text{факт}}$.

При вычислении $s^2_{\text{факт}}$ по формуле (5.6а) в роли случайных переменных выступают k групповых средних $\bar{x}_j$. На их значения накладывается одно ограничение (5.4а). Поэтому число степеней свободы факторной исправленной дисперсии равно:

$$\nu_{\text{факт}} = k - 1. \qquad (5.7)$$

В-третьих, можно использовать групповые исправленные дисперсии s_j^2. Из (5.2) следует, что сумма всех квадратов отклонений $\left(x_{ij} - \bar{x}_j\right)^2$ элементов j-й группы от их группового среднего равна $(n_j - 1) \cdot s_j^2$. Здесь $n_j - 1$ — число степеней свободы при вычислении s_j^2. С учетом (5.3) все k групповых исправленных дисперсий s_j^2 вместе дадут число степеней свободы, равное:

$$\nu_{\text{ост}} = N - k. \qquad (5.8)$$

Разделив сумму квадратов отклонений всех наблюдений x_{ij} от их групповых средних $\bar{x}_j$ на число степеней свободы $\nu_{\text{ост}}$, получим определение *остаточной* исправленной дисперсии:

$$s^2_{\text{ост}} = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2 + ... + (n_k - 1) \cdot s_k^2}{N - k}. \qquad (5.9)$$

Термин «остаточная» применяется потому, что оценка (5.9) не зависит от того, справедлива нулевая гипотеза H_0: $\mu_1 = \ldots = \mu_k$ или нет. Даже если групповые средние $\bar{x}_j$ различаются (правая колонка (d, e, f) на рис. 5.1), это никак не повлияет на разброс значений внутри каждой группы, потому что этот разброс определяется постоянной дисперсией σ^2 случайной величины X.

Сравнив (5.5), (5.7) и (5.8), легко заметить, что:

$$\nu_{\text{общ}} = \nu_{\text{факт}} + \nu_{\text{ост}}. \qquad (5.10а)$$

Можно показать, что общая, факторная и остаточная исправленные дисперсии связаны соотношением (без доказательства):

$$(N-1)\cdot s^2_{\text{общ}} = (k-1)\cdot s^2_{\text{факт}} + (N-k)\cdot s^2_{\text{ост}}. \qquad (5.10б)$$

Однократное сравнение $s^2_{\text{факт}}$ и $s^2_{\text{ост}}$ с помощью F-критерия Фишера–Снедекора позволяет проверить нулевую гипотезу H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. Таким образом, дисперсионный анализ заменяет собой многократное применение t-критерия Стьюдента для сравнения нескольких групповых средних $\bar{x}_j$. Поскольку необходимо провести лишь одно сравнение ($s^2_{\text{факт}}$ и $s^2_{\text{ост}}$), то использовать поправку Бонферрони (4.14) не требуется. Поэтому дисперсионный анализ — более мощный критерий сравнения нескольких средних $\bar{x}_j$, чем t-критерий Стьюдента.

Пример. Значение случайной величины X измеряли при трех различных уровнях фактора. Объемы групп измерений составили $n_1 = 20$, $n_2 = 30$ и $n_3 = 30$. Групповые средние оказались равны $\bar{x}_1 = 2$, $\bar{x}_2 = 7$ и $\bar{x}_3 = 5$. А исправленные групповые дисперсии $s_1^2 = 9$, $s_2^2 = 10$ и $s_3^2 = 8$. Вычислить общую, факторную и остаточную исправленные дисперсии.

Решение. Число уровней фактора $k = 3$.

Общий объем наблюдений (5.3): $N = 20 + 30 + 30 = 80$.

Общее среднее (5.4а): $\bar{x} = \dfrac{20\cdot 2 + 30\cdot 7 + 30\cdot 5}{80} = \dfrac{400}{80} = 5$.

Факторная исправленная дисперсия (5.6):

$$s^2_{\text{факт}} = \frac{20\cdot(2-5)^2 + 30\cdot(7-5)^2 + 30\cdot(5-5)^2}{3-1} = \frac{300}{2} = 150.$$

Остаточная исправленная дисперсия (5.9):

$$s^2_{\text{ост}} = \frac{(20-1)\cdot 9 + (30-1)\cdot 10 + (30-1)\cdot 8}{80-3} = \frac{171+290+232}{77} = \frac{693}{77} = 9.$$

Чтобы найти общую исправленную дисперсию, воспользуемся соотношением (5.10б): $(80-1)\cdot s^2_{\text{общ}} = (3-1)\cdot 150 + (80-3)\cdot 9$. Тогда:

$$s^2_{\text{общ}} = \frac{300+693}{79} = \frac{993}{79} = 12\frac{45}{79}.$$

Ответ: $s^2_{\text{факт}} = 150$, $s^2_{\text{ост}} = 9$, $s^2_{\text{общ}} = 12\dfrac{45}{79}$.

5.3. Однофакторный дисперсионный анализ, выявление влияния фактора

Однофакторный дисперсионный анализ предназначен для выявления влияния одного фактора на математическое ожидание $\mu(X)$ случайной величины X. Требуется, чтобы величина X была распределена нормально и

ее дисперсия σ^2 не изменялась под действием изучаемого фактора. Первичными данными для однофакторного дисперсионного анализа служат k групп (выборок) значений x_{ij} величины X, полученные при k различных уровнях фактора $A_1, A_2, \ldots, A_k$.

Нулевая гипотеза H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ предполагает равенство математических ожиданий μ_j случайной величины X при всех k уровнях фактора A_j. Когда эта гипотеза верна, факторная исправленная дисперсия $s^2_{\text{факт}}$, вычисленная по формуле (5.6), не должна значимо отличаться от остаточной исправленной дисперсии $s^2_{\text{ост}}$, вычисленной по формуле (5.9). Если же $\mu(X)$ существенно изменяется хотя бы при одном уровне фактора, то значение $s^2_{\text{факт}}$ увеличится, в то время как величина $s^2_{\text{ост}}$ значимо не изменится. Поэтому в числитель F-статистики берут $s^2_{\text{факт}}$, а в знаменатель отправляют $s^2_{\text{ост}}$:

$$F = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}}. \qquad (5.11)$$

Эта статистика используется для проверки основной гипотезы H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ против альтернативной гипотезы H_1, которая предполагает неравенство хотя бы одного μ_j всем остальным. (Если эта альтернативная гипотеза H_1 верна, то математическое ожидание $\mu(X)$ случайной величины X изменяется хотя бы при одном уровне фактора A_j.)

Критическая область статистики (5.11) правосторонняя: $F \geq F_{\text{кр}}$. Критические точки $F_{\text{кр}}(\alpha; k - 1; N - k)$ находят по распределению Фишера–Снедекора (прил. 5). Требуемый уровень значимости α. Число степеней свободы числителя равно $\nu_{\text{факт}} = k - 1$. Число степеней свободы знаменателя равно $\nu_{\text{ост}} = N - k$.

Уровень фактора А		
I	II	III
5	–2	14
19	–19	–9
24	–9	–2
6	5	–3
–1	–25	
7		

Пример. Значения случайной величины X измеряли при трех уровнях фактора A. Результаты измерений сведены в таблицу. Влияет ли фактор A на математическое ожидание случайной величины X? Требуется уровень значимости $\alpha = 0{,}05$. Считать, что величина X распределена нормально, а ее дисперсия σ^2 не зависит от фактора A.

Решение. Основная гипотеза H_0: $\mu_{\text{I}} = \mu_{\text{II}} = \mu_{\text{III}}$. Альтернативная гипотеза H_1 предполагает, что хотя бы одно математическое ожидание μ_j отличается от двух других. Проверим гипотезы с помощью однофакторного дисперсионного анализа.

Число уровней фактора: $k = 3$. Объемы групп значений, показанных в таблице, равны: $n_{\text{I}} = 6$, $n_{\text{II}} = 5$ и $n_{\text{III}} = 4$. Общий объем равен их сумме: $N = 6 + 5 + 4 = 15$.

Вычислим групповые средние: $\bar{x}_{\text{I}} = \frac{5+19+24+6-1+7}{6} = 10,$ $\bar{x}_{\text{II}} = \frac{-2-19-9+5-25}{5} = -10,$ $\bar{x}_{\text{III}} = \frac{14-9-2-3}{4} = 0$ и общее среднее: $\bar{x}_{\text{III}} = \frac{10 \cdot 6 - 10 \cdot 5 + 0 \cdot 4}{15} = \frac{10}{15} = \frac{2}{3}.$

Затем находим групповые исправленные дисперсии:

$$s_{\text{I}}^2 = \frac{(5-10)^2 + (19-10)^2 + (24-10)^2 + (6-10)^2 + (-1-10)^2 + (7-10)^2}{6-1} = \frac{448}{5} = 89{,}6.$$

$$s_{\text{II}}^2 = \frac{(-2+10)^2 + (-19+10)^2 + (-9+10)^2 + (5+10)^2 + (-25+10)^2}{5-1} = \frac{596}{5} = 149.$$

$$s_{\text{III}}^2 = \frac{(14-0)^2 + (-9-0)^2 + (-2-0)^2 + (-3-0)^2}{4-1} = \frac{196+81+4+9}{3} = \frac{290}{3} = 96\frac{2}{3}.$$

Теперь определим значение факторной исправленной дисперсии (5.6): $s_{\text{факт}}^2 = \frac{n_{\text{I}} \cdot (\bar{x}_{\text{I}} - \bar{x})^2 + n_{\text{II}} \cdot (\bar{x}_{\text{II}} - \bar{x})^2 + n_{\text{III}} \cdot (\bar{x}_{\text{III}} - \bar{x})^2}{k-1} =$

$$= \frac{6 \cdot \left(10 - \frac{2}{3}\right)^2 + 5 \cdot \left(-10 - \frac{2}{3}\right)^2 + 4 \cdot \left(0 - \frac{2}{3}\right)^2}{3-1} = \frac{1640}{3} = 546\frac{2}{3}.$$

Ее число степеней свободы (5.7): $\nu_{\text{факт}} = k - 1 = 3 - 1 = 2$.

Остаточная дисперсия (5.9) равна:

$$s_{\text{ост}}^2 = \frac{(n_{\text{I}} - 1) \cdot s_{\text{I}}^2 + (n_{\text{II}} - 1) \cdot s_{\text{II}}^2 + (n_{\text{III}} - 1) \cdot s_{\text{III}}^2}{N - k} =$$

$$= \frac{(6-1) \cdot \frac{448}{5} + (5-1) \cdot \frac{596}{4} + (4-1) \cdot \frac{290}{3}}{15-3} = \frac{1334}{12} = 111\frac{1}{6}.$$

Ее число степеней свободы (5.8): $\nu_{\text{ост}} = \text{N} - k = 15 - 3 = 12$.

Сравним значения $s_{\text{факт}}^2$ и $s_{\text{ост}}^2$ при помощи F-статистики (5.11): $F = \frac{s_{\text{факт}}^2}{s_{\text{ост}}^2} = \frac{1640}{3} : \frac{1334}{12} = \frac{1640 \cdot 4}{1334} \approx 4{,}9.$ Правосторонняя критическая точка распределения Фишера–Снедекора при требуемом уровне значимости $\alpha = 0{,}05$, числах степеней свободы числителя $\nu_{\text{факт}} = 2$ и знаменателя $\nu_{\text{ост}} = 12$ равна (прил. 5) $F_{\text{кр}}(0{,}05; 2; 12) \approx 3{,}8853$. Отвергаем нулевую гипотезу, поскольку значение статистики $F \approx 4{,}9$ попадает в критическую область $F \geq F_{\text{кр}}$.

***Ответ*:** да, фактор влияет на $\mu(X)$. (Принята H_1.)

5.4. Ограничения метода: нормальность распределения, гомогенность дисперсии

Выводы, полученные при помощи дисперсионного анализа, будут корректными только тогда, когда соблюдаются оба следующих ограничения.

Во-первых, все группы (выборки) должны представлять нормально распределенные статистические совокупности. Другими словами, распределение случайной величины *X* должно оставаться нормальным при всех уровнях фактора. Пусть, например, мы хотим сравнить математические ожидания массы таблетки в нескольких больших партиях. Тогда первое условие применимости дисперсионного анализа требует, чтобы распределение массы таблетки в каждой из этих партий было нормальным.

Это ограничение обусловлено применением F-статистики (5.11): $F = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}}$. Она подчиняется распределению Фишера–Снедекора (пункт 4.3), если и ее числитель, и знаменатель являются состоятельными и несмещенными оценками дисперсии нормально распределенных случайных величин. А поскольку и $s^2_{\text{факт}}$, и $s^2_{\text{ост}}$ вычисляются с использованием всех групп, то каждая группа должна представлять нормальное распределение.

Во-вторых, дисперсия σ^2 случайной величины *X* должна быть одинаковой (гомогенной) при всех уровнях фактора. (Дисперсия массы таблетки должна быть одинаковой во всех сравниваемых партиях.)

Если дисперсия σ^2 случайной величины *X* меняется от одного уровня фактора к другому, то ее следует оценивать отдельно в каждой группе (такими оценками будут s_j^2). А $s^2_{\text{факт}}$ и $s^2_{\text{ост}}$ потеряют свой смысл, поскольку они зависят от всех групп сразу. Напомним, что мы сравнивали их при помощи F-статистики (5.11) именно потому, что они выступали в роли двух статистических оценок одной и той же величины — независимой от фактора, одинаковой при всех его уровнях дисперсии σ^2 случайной величины *X*.

5.5. Понятие о двухфакторном и многофакторном дисперсионном анализе

Случайная величина *X* может подвергаться воздействию не одного, а сразу нескольких факторов, назовем их фактор *A*, фактор *B*, фактор *C* и т. д. Многофакторный дисперсионный анализ призван прояснить, влияют ли эти факторы на математическое ожидание (генеральное среднее) μ исследуемой случайной величины *X*. Ограничения многофакторного дисперсионного анализа те же, что у однофакторного. Во-первых, дисперсия σ^2 случайной величины *X* должна оставаться постоянной при любых ком-

бинациях уровней факторов. Во-вторых, распределение случайной величины X должно всегда оставаться нормальным.

Основная гипотеза H_0: $\mu(X) = \text{const}$ одинакова как для однофакторного, так и для многофакторного дисперсионного анализа. А вот альтернативную статистическую гипотезу H_1 можно сформулировать по-разному. Эта гипотеза может предполагать, что $\mu(X)$ изменяется под влиянием какого-то одного фактора, например фактора A. В этом случае группы значений величины X объединяются в соответствии с уровнями этого фактора A_j. Затем вычисляют факторную исправленную дисперсию $s^2_{\text{факт.}\,A}$ для фактора A и сравнивают ее с остаточной исправленной дисперсией $s^2_{\text{ост}}$ при помощи F-статистики (5.11), как и при однофакторном дисперсионном анализе. Аналогичным образом можно проверить влияние второго фактора B и всех остальных факторов.

Кроме отдельных факторов, эффект может оказывать их комбинация. Так, различные лекарственные средства могут взаимодействовать друг с другом. Одновременный прием нескольких взаимодействующих препаратов может дать побочный эффект, не вызываемый ни одним из этих средств по-отдельности. Каждой комбинации факторов соответствует своя факторная исправленная дисперсия. Паре взаимодействующих факторов AB соответствует $s^2_{\text{факт.}\,AB}$. Тройке факторов ABC соответствует $s^2_{\text{факт.}\,ABC}$ и т. д. Чтобы выяснить, влияет ли данная комбинация факторов на $\mu(X)$, соответствующую факторную исправленную дисперсию сравнивают с остаточной $s^2_{\text{ост}}$ при помощи F-статистики (5.11).

Используем двухфакторный дисперсионный анализ в простейшей ситуации, когда факторы A и B не взаимодействуют между собой, и при каждой комбинации этих факторов значение случайной величины X измерялось только один раз.

Пример. Таблица содержит значения случайной величины X, измеренные при $k = 3$ уровнях фактора A и $n = 3$ уровнях фактора B. Влияют ли эти факторы на $\mu(X)$? Требуемый уровень значимости $\alpha = 0{,}05$. Считать, что распределение X нормальное, а дисперсия постоянна: $\sigma^2(X) = \text{const}$.

		уровни фактора A		
		A_1	A_2	A_3
уровни фактора B	B_1	–5	–27	2
	B_2	1	4	13
	B_3	22	–1	–9

Решение. Основная гипотеза H_0: $\mu(X) = \text{const}$ предполагает, что ни один из факторов не влияет на математическое ожидание μ случайной величины X. Требуется проверить ее против двух альтернативных гипотез H_1: а) влияет фактор A; б) влияет фактор B.

Найдем средние значения.

По столбцам:

$$\bar{x}_{A1} = \frac{-5+1+22}{3} = 6,\ \bar{x}_{A2} = \frac{-27+4-1}{3} = -8 \text{ и } \bar{x}_{A3} = \frac{2+13-9}{3} = 2.$$

По строкам:

$$\bar{x}_{B1} = \frac{-5-27+2}{3} = -10,\ \bar{x}_{B2} = \frac{1+4+13}{3} = 6 \text{ и } \bar{x}_{B3} = \frac{22-1-9}{3} = 4.$$

Общее среднее: $\bar{x} = \frac{6-8+2}{3} = \frac{-10+6+4}{3} = 0.$

Теперь вычислим факторные исправленные дисперсии (5.6). Обусловленная фактором A: $s^2_{\text{факт. }A} = n \cdot \frac{(\bar{x}_{A1}-\bar{x})^2 + (\bar{x}_{A2}-\bar{x})^2 + (\bar{x}_{A3}-\bar{x})^2}{k-1} =$

$$= 3 \cdot \frac{(6-0)^2 + (-8-0)^2 + (2-0)^2}{3-1} = 3 \cdot \frac{104}{2} = 156.$$

Ее число степеней свободы (5.7): $\nu_{\text{факт. }A} = k - 1 = 3 - 1 = 2$.

Чтобы найти $s^2_{\text{факт. }B}$, обусловленную фактором B, надо от столбцов таблицы перейти к ее строкам:

$$s^2_{\text{факт. }B} = k \cdot \frac{(\bar{x}_{B1}-\bar{x})^2 + (\bar{x}_{B2}-\bar{x})^2 + (\bar{x}_{B3}-\bar{x})^2}{n-1} =$$

$$= 3 \cdot \frac{(-10-0)^2 + (6-0)^2 + (4-0)^2}{3-1} = 3 \cdot \frac{152}{2} = 228.$$

Ее число степеней свободы: $\nu_{\text{факт. }B} = n - 1 = 3 - 1 = 2$.

Поскольку все группы в таблице содержат только по одному значению, невозможно вычислить ни остаточную исправленную дисперсию $s^2_{\text{ост}}$, ни ее число степеней свободы $\nu_{\text{ост}}$, прибегнув к их определениям (5.9) и (5.8). Вместо этого можно использовать свойства (5.10), но теперь вместо одной $s^2_{\text{факт}}$ следует учитывать как $s^2_{\text{факт. }A}$, так и $s^2_{\text{факт. }B}$. Поэтому формула (5.10а) преобразуется к виду:

$$\nu_{\text{общ}} = \nu_{\text{факт. }A} + \nu_{\text{факт. }B} + \nu_{\text{ост}}.$$

А свойство (5.10б) с учетом двух факторных исправленных дисперсий:

$$\nu_{\text{общ}} \cdot s^2_{\text{общ}} = \nu_{\text{факт. }A} \cdot s^2_{\text{факт. }A} + \nu_{\text{факт. }B} \cdot s^2_{\text{факт. }B} + \nu_{\text{ост}} \cdot s^2_{\text{ост}}.$$

Найдем общую исправленную дисперсию (5.4б):

$$s^2_{\text{общ}} = \frac{(-5-0)^2 + (-27-0)^2 + (2-0)^2 + (1-0)^2 + (4-0)^2 + (13-0)^2 + (22-0)^2 + (-1-0)^2 + (-9-0)^2}{3 \cdot 3 - 1} =$$

$$= \frac{5^2 + 27^2 + 2^2 + 1^2 + 4^2 + 13^2 + 22^2 + 1^2 + 9^2}{9-1} = \frac{1510}{8}.$$

Ее число степеней свободы (5.5): $\nu_{\text{общ}} = N - 1 = n \cdot k - 1 = 3 \cdot 3 - 1 = 8$.

Теперь можно рассчитать число степеней свободы $\nu_{\text{ост}}$:

$$\nu_{\text{ост}} = \nu_{\text{общ}} - \nu_{\text{факт. }A} - \nu_{\text{факт. }B} = 8 - 2 - 2 = 4.$$

И саму остаточную исправленную дисперсию:

$$s^2_{\text{ост}} = \frac{\nu_{\text{общ}} \cdot s^2_{\text{общ}} - \nu_{\text{факт. }A} \cdot s^2_{\text{факт. }A} - \nu_{\text{факт. }B} \cdot s^2_{\text{факт. }B}}{\nu_{\text{ост}}} =$$

$$= \frac{8 \cdot \frac{1510}{8} - 2 \cdot 156 - 2 \cdot 228}{4} = \frac{742}{4} = 185{,}5.$$

Рассчитаем значения F-статистики, соответствующие обоим вариантам альтернативной гипотезы: $F_A = \frac{s^2_{\text{факт. }A}}{s^2_{\text{ост}}} = \frac{156}{185{,}5} \approx 0{,}84$ и $F_B = \frac{s^2_{\text{факт. }B}}{s^2_{\text{ост}}} = \frac{228}{185{,}5} \approx 1{,}23$. Ни одно из них не попадает в критическую область $F \geq F_{\text{кр}}$, где $F_{\text{кр}}(0{,}05;\ 2;\ 8) = 6{,}94$ — правосторонняя критическая точка распределения Фишера–Снедекора.

***Ответ*:** нет, ни фактор A, ни фактор B не влияют на $\mu(X)$.

6. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

6.1. Стохастическая и функциональная зависимости, корреляция

В теории вероятностей случайные события могут классифицироваться как независимые или зависимые. Напомним, что случайные события (не)зависимы, если вероятность одного из них (не)зависит от осуществления других. Распространим понятия зависимости и независимости на случайные величины:

1. Случайные величины X и Y *независимы*, если закон распределения одной из них не зависит от значения, которое принимает другая случайная величина.

2. Случайные величины *зависимы*, если их распределения зависят от того, какие значения принимают другие величины.

Такая зависимость случайных величин называется *стохастической*, или *статистической*. Если случайная величина Y зависит от величины X стохастически, то каждому возможному значению величины X соответствует определенный закон распределения Y. Такое распределение случайной величины Y называется *условным*, поскольку оно имеет место только при данном значении x. Условное распределение величины Y, справедливое при условии $X = x$, описывается *условной интегральной функцией* $F_Y(y/x)$. Для непрерывных условных распределений можно ввести *условную плотность вероятности*, например $f_Y(y/x)$.

Если X и Y — зависимые случайные величины, то они обе могут иметь условные распределения $F_Y(y/x)$ и $F_X(x/y)$.

Функциональная зависимость представляет собой предельный случай стохастической зависимости. Если Y является функцией X, то каждому значению x соответствует только одно определенное значение y. Фактически, условное распределение величины Y имеет только одно возможное значение y, условная вероятность которого $P(y/x) = 1$. (Можно сказать, что условное распределение $f_Y(y/x)$ «стягивается» в одну точку y.) Таким образом, функция создает наиболее сильную, абсолютную связь между величинами.

Условные распределения статистически зависимых случайных величин имеют свои характеристики — условные математические ожидания, условные дисперсии и т. д. Особый практический интерес вызывает частный случай стохастической зависимости, когда каждому значению одной величины, скажем $X = x$, отвечает определенное значение условного математического ожидания $\mu(Y/x)$ другой величины Y. Такая зависимость называется корреляционной зависимостью, или *корреляцией*. Корреляция — наиболее сильная разновидность стохастической зависимости, она наиболее близка к функции*.

Двум (или более) случайным величинам можно поставить в соответствие их совместное распределение. Интегральная функция $F(x; y)$ совместного распределения в точке с координатами $(x; y)$ равна вероятности того, что величина X будет меньше текущего значения x и одновременно величина Y будет меньше, чем y: $F(x; y) = P[(X < x) \cdot (Y < y)]$.

Совместная плотность вероятности $f(X; Y)$ двух непрерывных случайных величин X и Y определена на множестве точек плоскости $X0Y$. Она равна отношению вероятности ΔP одновременного попадания величин X и Y в бесконечно малые окрестности Δx и Δy точки $(x; y)$ к Δx и Δy:**

$$f(X;Y) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{\Delta P}{\Delta x \cdot \Delta y}. \qquad (6.1)$$

Совместную плотность вероятности двух случайных величин можно отобразить графически при помощи корреляционного поля, как показано на рис. 6.1. Каждая точка плоскости $X0Y$ закрашивается тем интенсивнее, чем выше в ней значение $f(X; Y)$. Абсолютно белыми остаются те области, вероятность попадания в которые равна нулю.

---

* Не следует отождествлять математические зависимости и причинно-следственные связи, наблюдаемые в окружающем нас мире. Например, статистическими методами можно доказать корреляцию между шумом деревьев и силой ветра. Однако вопрос о причине и следствии (шумят ли деревья так, что ветер поднялся, или наоборот, ветер усилился, и поэтому деревья зашумели?) выходит за рамки математики.

** Совместная плотность вероятности $f(X; Y)$ равна второй смешанной частной производной интегральной функции совместного распределения: $f(x; y) = \dfrac{\partial^2 F(x; y)}{\partial x \cdot \partial y}$.

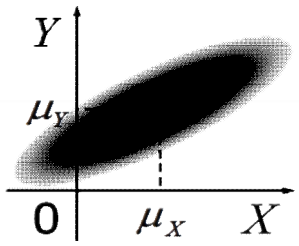

Рис. 6.1. Корреляционное поле двух случайных величин X и Y. Степень затемнения пропорциональна плотности совместного распределения вероятностей $f(X; Y)$. Математические ожидания случайных величин X и Y обозначены μ_X и μ_Y

6.2. Линейная регрессия. Коэффициент корреляции (Пирсона), его свойства, связь с параметрами линейной регрессии. Нелинейная регрессия

В математической статистике *регрессией* называется зависимость условного математического ожидания одной случайной величины от значений, принимаемых другими величинами. Например, зависимость условного математического ожидания $\mu(Y/x)$ случайной величины Y от x именуется регрессией случайной величины Y на X. Каждому возможному значению x отвечает единственное значение условного математического ожидания $\mu(Y/x)$, поэтому регрессия является функцией.

Зависимость $\mu(X/y)$ от y называется регрессией X на Y.

В силу исторически сложившейся традиции график регрессии часто именуют *линией регрессии*.

Если линия регрессии прямая, то регрессия называется *линейной*. Уравнение линейной регрессии Y на X можно записать в виде:

$$\mu(Y/x) = a \cdot x + b, \qquad (6.2а)$$

где a и b — постоянные коэффициенты этой регрессии. Аналогично, уравнение линейной регрессии X на Y с постоянными коэффициентами c и d:

$$\mu(X/y) = c \cdot y + d. \qquad (6.2б)$$

Модель линейной регрессии получила широкое распространение в статистике благодаря своей математической простоте. С другой стороны, если случайные величины X и Y имеют нормальные распределения, то их регрессии друг на друга будут линейными, как показано на рис. 6.2.

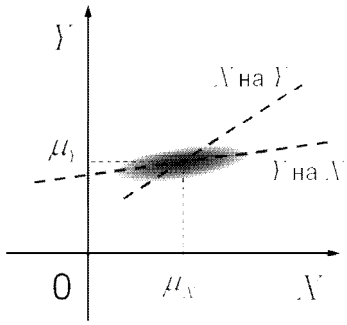

Рис. 6.2. Корреляционное поле двух нормально распределенных случайных величин X и Y. Штриховые линии показывают линейные регрессии Y на X и X на Y. Они пересекаются в центре корреляционного поля, координаты которого (μ_X; μ_Y)

Пусть случайная величина X распределена нормально, с дисперсией σ_X^2 и математическим ожиданием μ_X. И случайная величина Y также распределена нормально, с параметрами μ_Y и σ_Y^2. Тогда корреляцию между ними характеризует *коэффициент корреляции* (*Пирсона*):

$$\rho = \frac{\mu[(X-\mu_X)\cdot(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y}. \qquad (6.3)$$

Этот коэффициент корреляции обладает следующими свойствами:

1) Может принимать значения в интервале от –1 до 1:

$$|\rho| \le 1. \qquad (6.4)$$

Чем больше абсолютное значение ρ, тем теснее линейная корреляция, тем более вытянуто корреляционное поле. Например, линейная корреляция, представленная рис. 6.3, *d*, менее тесная, чем на рис. 6.3, *e* и 6.3, *f*.

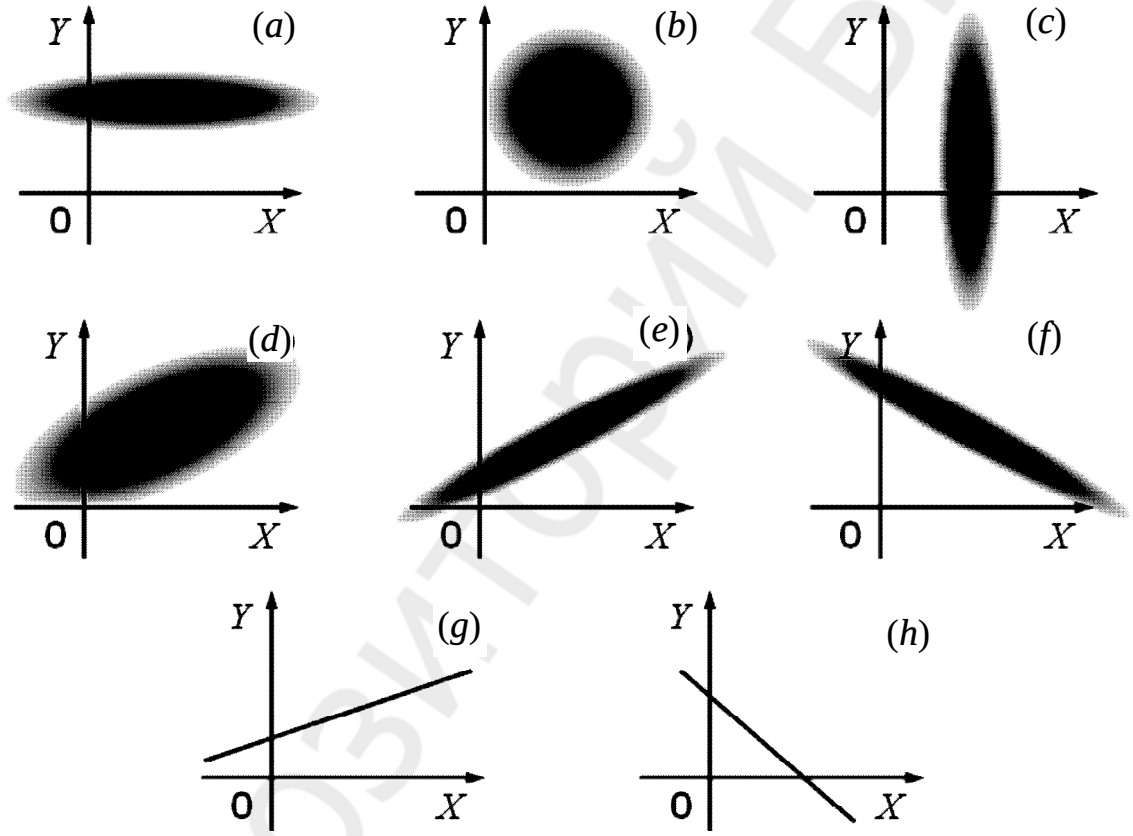

Рис. 6.3. Корреляционные поля нормально распределенных случайных величин X и Y. На (*a*), (*b*) и (*c*) величины X и Y независимы, коэффициент корреляции ρ = 0. На (*d*) и (*e*) корреляция положительная, ρ > 0. На (*f*) корреляция отрицательная, ρ < 0. На (*d*) корреляция слабее, чем на (*e*) и (*f*), $|\rho_d| < |\rho_e|$, $|\rho_d| < |\rho_f|$. На (*g*) и (*h*) корреляция выродилась в линейную функцию, $\rho_g = 1$ и $\rho_h = -1$

2) ρ = 0, когда нормально распределенные величины X и Y независимы. Корреляционное поле совместного распределения независимых нормальных величин не имеет наклона относительно координатных осей, как показано на рис. 6.3, *a*, 6.3, *b* и 6.3, *c* (оно или круглое, или вытянутое только вдоль оси 0X, или только вдоль 0Y).

3) $\rho = \pm 1$, если между X и Y существует функциональная линейная зависимость. В этом случае корреляционное поле «стягивается» в прямую линию, что показано на рис. 6.3, g и 6.3, h.

4) Тип корреляции. Линейная корреляция прямая, когда $\rho > 0$ (X и Y растут и убывают одновременно, как показано на рис. 6.3, d, 6.3, e и 6.3, g.) Если же $\rho < 0$, то линейная корреляция обратная. (X растет, когда Y убывает и наоборот. Обратная корреляция представлена рис. 6.3, f и 6.3, h.)

5) Коэффициент корреляции Пирсона (6.3) используется в уравнениях линейной регрессии (6.2) для нормально распределенных случайных величин:

$$\mu(Y/x) = \mu_Y + \rho \cdot \frac{\sigma_Y}{\sigma_X} \cdot (x - \mu_X), \qquad (6.5а)$$

$$\mu(X/y) = \mu_X + \rho \cdot \frac{\sigma_X}{\sigma_Y} \cdot (y - \mu_Y). \qquad (6.5б)$$

Линии регрессии (6.5) пересекаются в точке с координатами (μ_X; μ_Y), как показано на рис. 6.2. Эти линии совпадают при $\rho = \pm 1$, когда корреляция вырождается в линейную функцию.

Регрессия *нелинейная*, если ее график отличается от прямой линии. Уравнение нелинейной регрессии содержит более сложные функции, например степенные, логарифмические, показательные и т. д. При нелинейной регрессии корреляционное поле искривляется, что схематически обозначено на рис. 6.4.

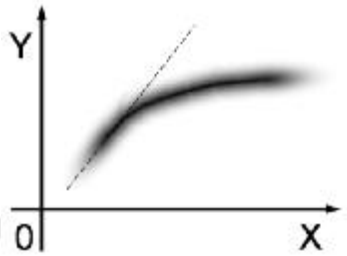

Рис. 6.4. Корреляционное поле искривлено при нелинейной регрессии

Следует особо отметить, что коэффициент корреляции Пирсона (6.3) не отражает тесноту связи при нелинейной регрессии. В некоторых случаях превратить нелинейную регрессию в линейную помогают математические преобразования величин. Подходящее преобразование величины Y в новую величину Z должно приводить уравнение регрессии к линейному виду $\mu(Z/x) = a \cdot x + b$ (с постоянными коэффициентами a и b).

Пример. Уравнение регрессии Y на X имеет вид: $\mu(Y/x) = a \cdot e^{bx}$. Коэффициенты a и b постоянные. Перейти к линейной регрессии.

Решение. Перейдем от Y к новой величине $Z = \ln Y$. Теперь уравнение регрессии стало линейным: $\mu(Z/x) = \ln a + b \cdot x$.

***Ответ*:** $\mu(Z/x) = \ln a + b \cdot x$, где $Z = \ln Y$.

6.3. Выборочный коэффициент корреляции. Проверка существенности корреляционной связи

Неизвестную связь между двумя случайными величинами X и Y можно оценить статистически. Для этого следует провести серию случайных экспериментов (наблюдений). В каждом эксперименте должны одновременно измеряться значения обеих величин X и Y. Например, чтобы оценить связь между длиной стебля растения (величиной X) и количеством содержащегося в нем биологически активного вещества (величиной Y), нужны несколько растений данного вида. У каждого растения следует измерить как длину стебля (значение x_i), так и содержание интересующего нас вещества (значение y_i). Тогда выборка объема n будет состоять из n пар значений $(x_i; y_i)$.

Графическим изображением такой выборки служит ее корреляционное поле — набор из n точек с координатами $(x_i; y_i)$, нанесенных на плоскость $X0Y$. Корреляционное поле выборки является статистической оценкой корреляционного поля генеральных совокупностей X и Y. Вероятность появления точек $(x_i; y_i)$ выше в тех местах плоскости $X0Y$, где больше совместная плотность вероятности $f(X; Y)$. Поэтому можно ожидать, что более темным областям генерального корреляционного поля (рис. 6.1–6.4) будут соответствовать более густые скопления точек на выборочном корреляционном поле.

Если корреляционное поле выборки визуально сгруппировано вдоль прямой линии (не искривлено), имеет смысл оценить тесноту возможной линейной связи между исследуемыми случайными величинами X и Y.

В определении коэффициента корреляции Пирсона (6.3) фигурируют неизвестные математические ожидания μ_X, μ_Y и неизвестные дисперсии σ_X^2, σ_Y^2 генеральных совокупностей X и Y. Заменим их соответствующими выборочными средними $\bar{x}$, $\bar{y}$ и исправленными выборочными дисперсиями s_X^2, s_Y^2. Получим *выборочный коэффициент корреляции* (Пирсона):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_X \cdot s_Y}. \qquad (6.6а)$$

Здесь вместо математического ожидания $\mu[(X - \mu_X) \cdot (Y - \mu_Y)]$, присутствующего в числителе формулы (6.3), используется его несмещенная оценка*:

* Деление на $n - 1$ вместо усреднения (деления на n) продиктовано переходом от математических ожиданий μ_X, μ_Y к выборочным средним $\bar{x}$, $\bar{y}$ (см. определение исправленной выборочной дисперсии (1.18)).

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{(n-1)}=\frac{(x_1-\bar{x})\cdot(y_1-\bar{y})+(x_2-\bar{x})\cdot(y_2-\bar{y})+\ldots+(x_n-\bar{x})\cdot(y_n-\bar{y})}{(n-1)}.$$

Определение (6.6а) можно преобразовать в более удобный для вычислений вид:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \qquad (6.6б)$$

Пример 1. Найти коэффициент корреляции Пирсона для выборки:

x_i	0	1	–1	0
y_i	5	4	3	4

Решение. Объем выборки $n = 4$.

Определим выборочные средние:

$$\bar{x}=\frac{0+1-1+0}{4}=0 \text{ и } \bar{y}=\frac{5+4+3+4}{4}=4.$$

Теперь можно вычислить выборочный коэффициент корреляции Пирсона по формуле (6.6б):

$$r=\frac{(0-0)\cdot(5-4)+(1-0)\cdot(4-4)+(-1-0)\cdot(3-4)+(0-0)\cdot(4-4)}{\sqrt{\left[(0-0)^2+(1-0)^2+(-1-0)^2+(0-0)^2\right]\cdot\left[(5-4)^2+(4-4)^2+(3-4)^2+(4-4)^2\right]}}=$$

$$=\frac{1}{\sqrt{2\cdot 2}}=0{,}5.$$

***Ответ*:** $r = 0{,}5$.

Даже при отсутствии корреляции между генеральными совокупностями X и Y (т. е. при $\rho = 0$) выборочный коэффициент корреляции Пирсона может отличаться от нуля ($r \neq 0$) под влиянием случайных факторов. Используем метод статистических гипотез. Основная гипотеза H_0: $\rho = 0$ предполагает отсутствие корреляции. Если же права альтернативная гипотеза H_1: $\rho \neq 0$, то корреляция имеется. Выяснить, которая из этих двух гипотез права в случае нормально распределенных величин X и Y, помогает Т-статистика (без доказательства):

$$T=r\cdot\sqrt{\frac{n-2}{1-r^2}}. \qquad (6.7)$$

При справедливой нулевой гипотезе H_0: $\rho = 0$ статистика (6.7) подчиняется распределению Стьюдента с числом степеней свободы:

$$\nu = n - 2. \qquad (6.8)$$

Критическая область $|t| \geq t_{кр}$ двусторонняя — в пользу существования корреляции между генеральными совокупностями X и Y говорят лишь достаточно большие абсолютные значения Т-статистики (6.7).

Пример 2. Свидетельствует ли о наличии какой-либо корреляции между случайными величинами X и Y выборочный коэффициент корреля-

ции Пирсона, найденный в примере 1? Требуемый уровень значимости $\alpha = 0{,}05$.

Решение. Напомним, что в примере 1 было вычислено значение коэффициента корреляции $r = 0{,}5$ для выборки объема $n = 4$.

Предположим, что случайные величины X и Y распределены нормально. Тогда нулевую гипотезу H_0: $\rho = 0$ можно проверить против альтернативной гипотезы H_1: $\rho \neq 0$ при помощи Т-статистики (6.7). Найдем ее значение:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} = 0{,}5 \cdot \sqrt{\frac{4-2}{1-0{,}5^2}} = \sqrt{\frac{2}{3}} \approx 0{,}82.$$

Число степеней свободы (6.8): $\nu = 4 - 2 = 2$. Двустороннюю критическую точку распределения Стьюдента при $\alpha = 0{,}05$ и $\nu = 2$ найдем в прил. 4: $t_{кр}(0{,}05;\ 2) \approx 4{,}3$. Вычисленное значение статистики $t \approx 0{,}82$ не попадает в критическую область $|t| \geq t_{кр}$, поэтому принимаем основную гипотезу H_0: $\rho = 0$.

***Ответ*:** нет, наличие корреляции не доказано при $\alpha = 0{,}05$. (Слишком мал объем выборки $n = 4$.)

Этот пример показал, что даже достаточно большие, на первый взгляд, абсолютные значения выборочного коэффициента корреляции (6.6) не всегда говорят о наличии какой-либо (даже самой слабой) корреляции между исследуемыми генеральными совокупностями. Выразим из формулы (6.7) значение $r_{кр}$, которое соответствует критическому значению Т-статистики: $t_{кр} = r_{кр} \cdot \sqrt{\frac{n-2}{1-r_{кр}^2}}$. Возведя это равенство в квадрат и затем приведя подобные слагаемые, получим:

$$|r_{кр}| = \frac{t_{кр}}{\sqrt{t_{кр}^2 + n - 2}}. \qquad (6.9)$$

Критические значения (6.9) выборочного коэффициента корреляции Пирсона зависят от объема выборки n и от требуемого уровня значимости α (через $t_{кр}(\alpha;\ n-2)$). О существовании корреляции можно говорить только тогда, когда абсолютное значение r превышает $|r_{кр}|$. Значения $|r_{кр}|$, вычисленные при уровне значимости $\alpha = 0{,}05$ для некоторых значений объема выборки n, показаны в табл. 6.1. Видно, что чем больше выборка, тем проще судить о предполагаемой корреляции исследуемых случайных величин.

Таблица 6.1

Критические значения выборочного коэффициента корреляции Пирсона при объеме выборки n и уровне значимости $\alpha = 0{,}05$

n	3	4	5	10	30	100	300	1000
$\lvert r_{кр}\rvert$	0,9969	0,9500	0,8783	0,6319	0,3610	0,1966	0,1133	0,0620

6.4. Оценка параметров линейной регрессии по данным выборки

Уравнения линейной регрессии (6.5) содержат неизвестные параметры случайных величин X и Y. Это среднеквадратичные отклонения σ_X и σ_Y, математические ожидания μ_X и μ_Y, коэффициент корреляции ρ генеральных совокупностей X и Y. Мы располагаем лишь выборкой, которая состоит из n пар значений $(x_i; y_i)$ этих случайных величин. Наилучшими статистическими оценками интересующих нас параметров генеральных совокупностей являются выборочные исправленные среднеквадратичные отклонения s_X и s_Y, выборочные средние $\bar{x}$ и $\bar{y}$, выборочный коэффициент корреляции r. Подставив их в уравнения (6.5), получим выборочные уравнения регрессии:

$$\bar{Y}/x = \bar{y} + r \cdot \frac{s_Y}{s_X} \cdot (x - \bar{x}). \qquad (6.10а)$$

$$\bar{X}/y = \bar{x} + r \cdot \frac{s_X}{s_Y} \cdot (y - \bar{y}). \qquad (6.10б)$$

Уравнение (6.10а) описывает линейную зависимость условного среднего $\bar{Y}/x$ от значения x. Эта зависимость называется выборочной регрессией Y на X. Уравнение (6.10б) описывает выборочную регрессию X на Y. Графики уравнений (6.10) называют выборочными линиями регрессии. Эти линии пересекаются в точке с координатами $(\bar{x}, \bar{y})$. Данная точка называется центром рассеяния, поскольку остальные точки корреляционного поля выборки «рассеяны» вокруг нее, как показано на рис. 6.5.

Угол между выборочными линиями регрессии зависит от абсолютной величины выборочного коэффициента корреляции. При $r = \pm 1$ обе эти линии сливаются в одну прямую, на которой лежат все точки выборки. А при $r = 0$ выборочные линии регрессии идут параллельно координатным осям $0X$ и $0Y$, пересекаясь под прямым углом.

Пример. Построить корреляционное поле и линии регрессии выборки:

x_i	9	–4	13	–6	8	–12	9	–5	–16	4
y_i	13	12	18	–3	22	–19	–7	–9	–26	–1

(6.11)

Решение. Корреляционное поле этой выборки содержит $n = 10$ точек, как показано на рис. 6.5.

Вычислим параметры выборки, используемые в выборочных уравнениях регрессии (6.10). Выборочные средние случайных величин X и Y:

$$\bar{x} = \frac{9-4+13-6+8-12+9-5-16+4}{10} = 0 \text{ и}$$

$$\bar{y} = \frac{13+12+18-3+22-19-7-9-26-1}{10} = 0.$$

Выборочные исправленные среднеквадратичные отклонения:

$$s_X = \sqrt{\frac{9^2 + (-4)^2 + 13^2 + (-6)^2 + 8^2 + (-12)^2 + 9^2 + (-5)^2 + (-16)^2 + 4^2}{10-1}} =$$

$$= \frac{\sqrt{888}}{3} \approx 9{,}933;$$

$$s_Y = \sqrt{\frac{13^2 + 12^2 + 18^2 + (-3)^2 + 22^2 + (-19)^2 + (-7)^2 + (-9)^2 + (-26)^2 + (-1)^2}{10-1}} =$$

$$= \frac{\sqrt{2298}}{3} \approx 15{,}979.$$

Выборочный коэффициент корреляции Пирсона (6.6):

$$r = \frac{9 \cdot 13 + (-4) \cdot 12 + 13 \cdot 18 + 6 \cdot 3 + 8 \cdot 22 + 12 \cdot 19 - 9 \cdot 7 + 5 \cdot 9 + 16 \cdot 26 - 4 \cdot 1}{\sqrt{888 \cdot 2298}} =$$

$$= \frac{1119}{\sqrt{2040624}} \approx 0{,}783.$$

Теперь можно переходить к выборочным уравнениям регрессии. Выборочная регрессия Y на X (6.10а):

$$\overline{Y}/x = 0 + \frac{1119}{\sqrt{888 \cdot 2298}} \cdot \frac{\frac{\sqrt{2298}}{3}}{\frac{\sqrt{888}}{3}} \cdot (x - 0) = \frac{1119}{888} \cdot x = 1\frac{77}{296} \cdot x \approx 1{,}26 \cdot x.$$

Выборочная регрессия X на Y (6.10б):

$$\overline{X}/y = 0 + \frac{1119}{\sqrt{888 \cdot 2298}} \cdot \frac{\frac{\sqrt{888}}{3}}{\frac{\sqrt{2298}}{3}} \cdot (y - 0) = \frac{1119}{2298} \cdot y = \frac{373}{766} \cdot y \approx 0{,}487 \cdot y.$$

Эти прямые пересекаются в точке с координатами (0; 0), как показано на рис. 6.5.

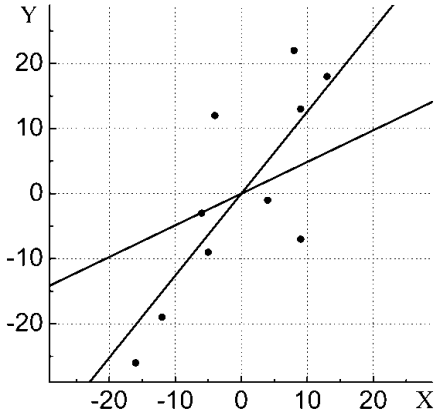

Рис. 6.5. Корреляционное поле выборки (6.11). Выборочный коэффициент корреляции Пирсона r = 0,783. Выборочные линии регрессии: $\overline{Y}/x = 1{,}26 \cdot x$ и $\overline{X}/y = 0{,}487 \cdot y$

Замечание. Выборочная линейная регрессия (6.10) является статистической оценкой линейной модели (6.5), которая справедлива в случае двух нормально распределенных генеральных совокупностей X и Y. Однако распределения случайных величин X и Y не всегда нормальны, а регрессия не всегда линейна. Первичная проверка применимости линейной модели заключается в построении корреляционного поля выборки. Это поле не должно иметь явных изгибов (как на рис. 6.4), точки «рассеяны» по сторонам воображаемой прямой линии.

Если объем выборки достаточно велик, можно построить оценку произвольной, даже нелинейной, регрессии. Рассмотрим, для определенности, регрессию величины Y на X. Разобьем весь диапазон попавших в выборку значений величины X на k частичных интервалов. Сгруппируем элементы выборки $(x_i; y_i)$ по этим интервалам. Для каждого интервала вычислим групповое среднее $\overline{y}_j$ значений величины Y, попавших в данный интервал. Определим середины интервалов x_j. Нанесем на плоскости $X0Y$ точки с координатами $(x_j; \overline{y}_j)$ и соединим их линиями, как на рис. 6.6. Получившаяся оценка свободна от каких-либо предположений о виде регрессии Y на X.

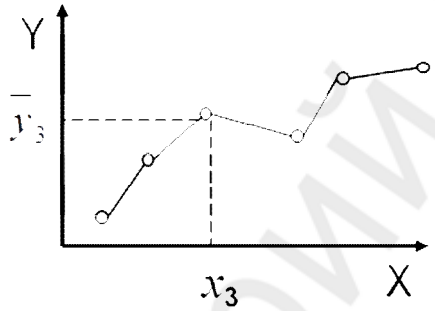

Рис. 6.6. Сгруппированная оценка линии регрессии Y на X. Отмечены середина 3-го интервала x_3 и групповое среднее $\overline{y}_3$ элементов выборки, попавших в этот интервал

6.5. Непараметрический коэффициент корреляции (Спирмена)

На практике встречаются величины, значения которых нельзя выразить количественно, численно. Эти значения можно упорядочить по их качеству, используя соотношение «лучше – хуже». На первое место поставим значение с самым высоким качеством, затем расположим остальные значения в порядке убывания их качества. Номера значений, упорядоченных по их качеству, на математическом языке называются рангами этих значений. Если качество нескольких значений одинаковое, то всем им приписывается одинаковый, средний ранг. (Например, оба значения, разделившие 1-е и 2-е места, получают ранг, равный 1,5.)

Между качественными величинами, как и между количественными, может существовать корреляция. Напомним, что теснота линейной корреляции между количественными нормально распределенными случайными величинами характеризуется коэффициентом корреляции Пирсона ρ, определение которого дает формула (6.3). Его статистическая оценка — выборочный коэффициент корреляции Пирсона r, рассчитываемый по

формулам (6.6). Если случайные величины X и Y качественные, то известны лишь ранги $\mathrm{Rank}X_i$ и $\mathrm{Rank}Y_i$ их значений, попавших в выборку. По аналогии с формулами (6.6) можно определить выборочный ранговый коэффициент корреляции Спирмена:

$$r_s = \frac{\sum_{i=1}^{n}(\mathrm{Rank}X_i - \overline{\mathrm{Rank}X}) \cdot (\mathrm{Rank}Y_i - \overline{\mathrm{Rank}Y})}{\sqrt{\sum_{i=1}^{n}(\mathrm{Rank}X_i - \overline{\mathrm{Rank}X})^2 \cdot \sum_{i=1}^{n}(\mathrm{Rank}Y_i - \overline{\mathrm{Rank}Y})^2}}. \quad (6.12)$$

Здесь использованы обозначения $\overline{\mathrm{Rank}X}$ и $\overline{\mathrm{Rank}Y}$ средних рангов значений случайных величин X и Y, попавших в выборку объема n.

Если в выборке нет повторений, все значения x_i имеют разные ранги $\mathrm{Rank}X_i$ и все значения y_i имеют разные ранги $\mathrm{Rank}Y_i$. В этом случае определение выборочного рангового коэффициента корреляции Спирмена (6.12) можно переписать, используя разницы рангов $\mathrm{Rank}X_i$ и $\mathrm{Rank}Y_i$ для каждого элемента выборки (без доказательства):

$$r_s = 1 - 6 \cdot \frac{\sum_{i=1}^{n}(\mathrm{Rank}X_i - \mathrm{Rank}Y_i)^2}{n \cdot (n^2 - 1)}. \quad (6.13)$$

Когда все точки корреляционного поля выборки лежат на монотонно возрастающей линии, ранги $\mathrm{Rank}X_i$ и $\mathrm{Rank}Y_i$ совпадают для каждого элемента выборки. При этом $\mathrm{Rank}X_i - \mathrm{Rank}Y_i = 0$ и, следовательно, $r_s = 1$. Если же Y является монотонно убывающей функцией величины X, то $\mathrm{Rank}Y_i = -\mathrm{Rank}X_i$. Можно показать, что в этом случае $r_s = -1$. Чем слабее корреляция между случайными величинами X и Y, тем ближе к нулю абсолютное значение r_s. Коэффициент корреляции Спирмена, рассчитанный по всем возможным значениям статистически независимых генеральных совокупностей X и Y, будет равен нулю: $\rho_s = 0$.

В отличие от коэффициента корреляции Пирсона, коэффициент корреляции Спирмена можно использовать, когда:

– исследуемые признаки (случайные величины X и Y) качественные;

– их распределения не подчиняются нормальному закону;

– связь между X и Y нелинейная.

Проверить основную гипотезу H_0: $\rho_s = 0$ против альтернативной H_1: $\rho_s \neq 0$ можно, вычислив значение статистики:

$$T_s = r_s \cdot \sqrt{\frac{n-2}{1-r_s^2}}. \quad (6.14)$$

Наличие связи между генеральными совокупностями X и Y предполагает альтернативная гипотеза H_1: $\rho_s \neq 0$. Ее следует признать справедливой, когда статистика (6.14) попадает в двустороннюю критическую область $|t_s| > t_s^{\text{кр}}$.

Следует учитывать, что в отличие от Т-статистики (6.7), применяемой для проверки значимости выборочного коэффициента корреляции Пирсона, статистика T_s подчиняется распределению Стьюдента с $\nu = n - 2$ степенями свободы лишь при большом ($n \to \infty$) объеме выборки.

Таблица 6.2

Критические значения выборочного коэффициента корреляции Спирмена при объеме выборки n и уровне значимости $\alpha = 0{,}05$

(Статистика в медицине и биологии: рук. / под ред. Ю. М. Комарова. М. : Медицина, 2000. Т. 1. Теоретическая статистика)

n	6	7	8	9	10	20	50
$\lvert r_s^{\text{кр}} \rvert$	0,886	0,786	0,738	0,7	0,648	0,447	0,279

Пример. Вычислить коэффициент корреляции Спирмена для выборки:

x_i	9	–4	–6	8	–12	–5	–16	4
y_i	13	12	–3	22	–19	–9	–26	–1

Решение. Всего в выборке $n = 8$ пар значений. Проранжируем значения x_i и y_i, дополним таблицу строками рангов RankX_i и RankY_i:

x_i	9	–4	–6	8	–12	–5	–16	4
RankX_i	8	5	3	7	2	4	1	6
y_i	13	12	–3	22	–19	–9	–26	–1
RankY_i	7	6	4	8	2	3	1	5

Поскольку значений с одинаковыми рангами нет, используем (6.13):

$$r_S = 1 - 6 \cdot \frac{(8-7)^2 + (5-6)^2 + (3-4)^2 + (7-8)^2 + (2-2)^2 + (4-3)^2 + (1-1)^2 + (6-5)^2}{8 \cdot (8^2 - 1)} =$$

$$= \frac{13}{14} \approx 0{,}93.$$

Ответ: $r_s = \frac{13}{14} \approx 0{,}93.$

6.6. Понятие о множественной корреляции

Распространены ситуации, когда случайные величины стохастически связаны сразу с несколькими другими величинами. Так, дневной объем продаж препарата аптекой связан с тем, какую аптеку из аптечной сети мы выберем, какой конкретно препарат имеется в виду, а также с интенсивностью рекламных кампаний, со временем года и т. д.

Ограничимся исследованием стохастической зависимости одной случайной величины Z от двух других — X и Y. Например, Z может обозначать физиологическую реакцию организма на введенную дозу X по прошествии времени Y с момента введения препарата. Величину Z называют

зависимой, а X и Y именуют предикторами, или объясняющими переменными. Зависимость условного математического ожидания $\mu(Z/x, y)$ от значений предикторов x и y описывается уравнением множественной регрессии. Наиболее распространена линейная модель множественной регрессии:

$$\mu(Z/x, y) = a_1 \cdot x + a_2 \cdot y + b. \qquad (6.15)$$

Чтобы статистически оценить корреляцию трех случайных величин Z, X и Y, необходима выборка, каждый элемент которой представляет собой тройку значений (x_i; y_i; z_i), полученных в одном эксперименте (наблюдении). Такая выборка позволяет рассчитать три выборочных коэффициента корреляции Пирсона r_{XZ}, r_{YZ} и r_{XY}. Каждый из них служит оценкой тесноты соответствующей парной линейной корреляции.

Тесноту корреляции между зависимой случайной величиной Z и парой ее предикторов X и Y описывает коэффициент множественной корреляции P. Если Z не зависит от X и Y, то $P = 0$. Если же Z является линейной функцией предикторов, то $P = 1$. Статистической оценкой P служит выборочный коэффициент множественной линейной корреляции:

$$R = \sqrt{\frac{r_{XZ}^2 - 2 \cdot r_{XZ} \cdot r_{XY} \cdot r_{YZ} + r_{YZ}^2}{1 - r_{XY}^2}}. \qquad (6.16)$$

Замечание. Наиболее надежны те статистические модели множественной регрессии, которые используют минимальное количество предикторов. Кроме того, чем слабее парные корреляции учитываемых предикторов, тем меньшими будут погрешности статистической модели.

7. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

7.1. Виды временных рядов и их характеристики. Тренд и случайная составляющая

Временной ряд — последовательность значений x_1, x_2, …, x_n величины X, полученных по мере возрастания времени. Практическое значение имеют *дискретные* временные ряды, значения которых разделены конечными интервалами времени. В дальнейшем будем полагать, что значения x_t соответствуют равноотстоящим моментам времени $t = 1, 2, \ldots, n$.

Изменение величины X во времени называется стохастическим (случайным) *процессом*. Этот термин напоминает, что в окружающем нас мире ничто не может измениться мгновенно*. Поэтому значение x_t связано с предыдущими значениями x_1, x_2, …, x_{t-1}. Этим обусловлено главное отличие временного ряда от выборки объема n — элементы репрезентативной выборки X_1, X_2, …, X_n должны быть независимы друг от друга.

* Законы физики утверждают, что никакое движение не может происходить со скоростью, большей, чем скорость света в вакууме.

Временные ряды могут быть получены двумя способами.

Пусть величина X непрерывно изменяется во времени. Так себя ведут, например, значения биохимических показателей, курс акций фармацевтических компаний, концентрация реагентов в рабочей зоне реактора. Значения x_t, измеренные в моменты времени $t = 1, 2, \ldots, n$, составляют *моментный* временной ряд.

Другие величины, например, объем реализации фармацевтического средства, число посетителей аптеки, количество новых препаратов, запущенных в производство, требуют накопления в течение интервалов времени (за год, за квартал, за день и т. д.). Так формируются *интервальные* временные ряды.

Если величина X зависит от времени функционально, то ее значения составляют *детерминированный* временной ряд. Математическую статистику больше интересуют *стохастические* (*случайные*) временные ряды. Распределение случайной величины X, изменяющейся в случайном процессе, зависит от времени t и от предыдущих значений $x_1, \ldots, x_{t-1}$. В каждый момент времени t это распределение имеет математическое ожидание $\mu(t) = \mu(x_t)$, дисперсию $\sigma_t^2 = \sigma^2(x_t)$ и другие параметры.

Элементы временного ряда x_t и x_{t+k} разделяет задержка (*временной лаг*) k. Тесноту линейной стохастической связи этих элементов ряда характеризует *автокорреляция* k-го порядка:

$$\rho_k = \frac{\mu[(x_t - \mu_t) \cdot (x_{t+k} - \mu_{t+k})]}{\sigma_t \cdot \sigma_{t+k}}. \qquad (7.1)$$

Будем считать, что стохастический временной ряд (значения x_t) состоит из двух компонент — детерминированной (y_t) и случайной (ε_t):

$$x_t = y_t + \varepsilon_t. \qquad (7.2)$$

Детерминированная составляющая y_t называется основной *тенденцией*, или *трендом*, временного ряда. При решении прикладных задач считается, что тренд изменяется медленно, его направление сохраняется в течение значительного промежутка времени. Наряду с общей тенденцией временного ряда могут наблюдаться периодически повторяющаяся *сезонная* составляющая и *циклы*. Например, продажи противовирусных препаратов могут расти год от года. На этот долговременный тренд могут накладываться сезонные колебания, обусловленные более высокой заболеваемостью населения в холодное время года. А изменения эпидемиологической обстановки («птичий грипп», «свиной грипп» и т. д.) могут вызывать циклические всплески продаж.

Наиболее прост статистический анализ *стационарных* случайных временных рядов, у которых:

1) $\mu_1 = \mu_2 = \ldots = \mu_n = \mu$ — математическое ожидание не зависит от времени;

2) $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$ — дисперсия не зависит от времени;

3) автокорреляция $\rho_k = \frac{\mu[(x_t - \mu)\cdot(x_{t+k} - \mu)]}{\sigma^2}$ одинакова для всех пар x_t и x_{t+k} элементов ряда, разделенных временным лагом k.

Если стационарный стохастический процесс не имеет сезонной или циклической составляющей (выполняется условие $\lim_{n\to\infty} \frac{1}{n}\cdot\sum_{k=1}^{n}\rho_k = 0$), то состоятельной и несмещенной оценкой его математического ожидания μ является среднее значение ряда:

$$\bar{x} = \frac{1}{n}\cdot\sum_{t=1}^{n} x_t. \qquad (7.3)$$

А состоятельной и несмещенной оценкой дисперсии σ^2 будет исправленная дисперсия:

$$s^2 = \frac{1}{n-1}\cdot\sum_{t=1}^{n}(x_t - \bar{x})^2. \qquad (7.4)$$

7.2. Сглаживание временных рядов. Определение линейного тренда ряда методом наименьших квадратов

Будем рассматривать временные ряды, случайная составляющая которых ε_t стационарна и, более того, ее математическое ожидание равно нулю: $\mu(\varepsilon_t) = 0$. Такой ряд получается, например, при измерении детерминированной величины X через равные промежутки времени при помощи измерительного прибора, если случайная погрешность ε всегда имеет один и тот же закон распределения. Каждое показание измерительного прибора x_t складывается из истинного значения измеряемой величины X и из ошибки измерения ε_t. Поскольку ошибка случайная (не систематическая), то $\mu(\varepsilon_t) = 0$. Постоянство дисперсии $\sigma^2(\varepsilon_t) = \sigma^2 = \text{const}$ означает, что разброс показаний относительно истинного значения измеряемой величины не меняется с течением времени:

$$x_t = X + \varepsilon_t. \qquad (7.5)$$

Зависимость от времени истинного значения величины X является трендом временного ряда (7.5). А случайная ошибка измерения представляет собой случайную компоненту ε_t этого ряда. Последовательность ε_1, ε_2, …, ε_n сама является стационарным случайным рядом с нулевым математическим ожиданием $\mu(\varepsilon_t) = 0$. Каким же образом оценить тренд ряда (7.5)? (Как определить истинную зависимость величины X от времени, если значения ряда (7.5) содержат случайные ошибки?)

Можно выделить два подхода к определению тренда временного ряда. Во-первых, если вид зависимости $X(t)$ уже известен и требуется найти только несколько постоянных коэффициентов (параметров этой функции времени), то применяют так называемый метод наименьших квадратов.

Во-вторых, если вид зависимости $X(t)$ неизвестен, то можно сглаживать временной ряд, усредняя его значения по нескольким соседним точкам.

Предположим, что вид функции $X(t)$ известен. По значениям ряда x_1, x_2, …, x_n требуется оценить параметры тренда a и b. Выразим из уравнения (7.5) случайную ошибку «измеренного» значения ряда: $\varepsilon_t = x_t - X$. Суть метода наименьших квадратов заключается в подборе таких параметров функции $X(t)$, при которых сумма квадратов случайных ошибок всех значений ряда минимальная:

$$SS = \sum_{t=1}^{n} (x_t - X) = \min. \qquad (7.6)$$

Найдем МНК-оценки параметров a и b линейной модели тренда:

$$X = a \cdot t + b. \qquad (7.7)$$

«Измеренному» в момент времени t значению ряда x_t соответствует «теоретическое» значение $a \cdot t + b$. Квадрат их разницы равен $(x_t - a \cdot t + b)^2$. Требуется минимизировать сумму квадратов:

$$SS = \sum_{t=1}^{n} (x_t - a \cdot t + b)^2.$$

В точке минимума производная функции равна нулю. Мы должны минимизировать сумму квадратов SS, используя параметры a и b. Для этого следует найти и приравнять к нулю частные производные $(SS)'_a$ и $(SS)'_b$. Проделав все необходимые выкладки, можно убедиться в том, что метод наименьших квадратов приводит к такому же уравнению линейного тренда, как уравнения выборочной линейной регрессии (6.10). Используем среднее значение ряда (7.3) и среднее время ряда* $\bar{t} = \dfrac{1 + 2 + \ldots + n}{n}$. Тогда МНК-уравнение линейного тренда (без доказательства):

$$X = \bar{x} + \frac{\sum_{t=1}^{n} \left(x_t - \bar{x}\right) \cdot \left(t - \bar{t}\right)}{\sum_{t=1}^{n} \left(t - \bar{t}\right)^2} \cdot \left(t - \bar{t}\right). \qquad (7.8)$$

Пример. Значения временного ряда (7.9) получены через единичные интервалы времени: $t = 1, 2, \ldots, n$. Используя метод наименьших квадратов, определить линейный тренд этого ряда. Изобразить ряд и его тренд на графике.

0 3 9 13 16 16 24 26 28 (7.9)

* Если время ряда «пробегает» все целые значения $t = 1, 2, \ldots, n$, то можно использовать соотношение $1 + 2 + \ldots + n = \dfrac{n \cdot (n+1)}{2}$. Тогда среднее время ряда $\bar{t} = \dfrac{n+1}{2}$.

Решение. Всего имеется $n = 9$ значений ряда.

Среднее значение ряда: $\bar{x} = \dfrac{0+3+9+13+16+16+24+26+28}{9} = 15$.

Среднее время ряда $\bar{t} = \dfrac{n+1}{2} = \dfrac{9+1}{2} = 5$.

Сумма в числителе уравнения линейного тренда (7.8):

$$\sum_{t=1}^{9}(x_t - 15)\cdot(t-5) = (0-15)\cdot(1-5) + (3-15)\cdot(2-5) + (9-15)\cdot(3-5) +$$
$$+ (13-15)\cdot(4-5) + (16-15)\cdot(5-5) + (16-15)\cdot(6-5) + (24-15)\cdot(7-5) +$$
$$+ (26-15)\cdot(8-5) + (28-15)\cdot(9-5) = 214.$$

Сумма в знаменателе уравнения линейного тренда (7.8):

$$\sum_{t=1}^{9}(t-5)^2 = (-4)^2 + (-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 + 4^2 = 60.$$

Подставляем найденные суммы в уравнение (7.8): $X = 15 + \dfrac{214}{60}\cdot(t-5)$. Перегруппировав слагаемые, получаем уравнение линейного тренда: $X = \dfrac{107\cdot t - 85}{30}$. Его график показан на рис. 7.1 прямой линией.

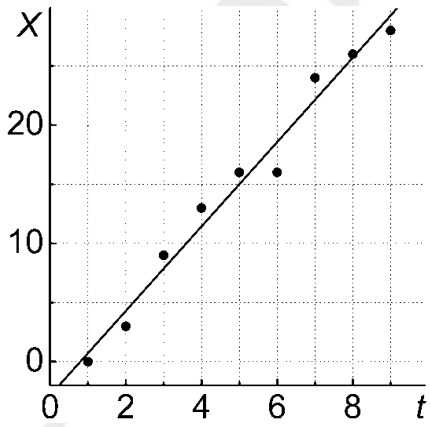

Рис. 7.1. Временной ряд (7.9) и его линейный тренд $X = \dfrac{107\cdot t - 85}{30}$

7.3. Сглаживание временных рядов методом скользящего среднего, экспоненциальное сглаживание

Будем полагать, что неизвестный тренд $X(t)$ ряда (7.5) изменяется достаточно медленно, и только случайная составляющая ε_t может значительно флуктуировать от наблюдения к наблюдению. Тогда на коротком временном интервале график тренда мало отличается от касательной прямой. Значит, в малой окрестности точки t, скажем, от $t - 1$ до $t + 1$, тренд можно считать практически линейным и оценивать его значение $X(t)$ по

методу наименьших квадратов. Среднее значение ряда в этой окрестности равно: $\bar{x} = \frac{x_{t-1} + x_t + x_{t+1}}{3}$, а среднее время равно: $\bar{t} = \frac{(t-1) + t + (t+1)}{3} = t$. Следовательно, в уравнении (7.8) множитель $\left(t - \bar{t}\right)$ равен нулю. А оценка x_t^* значения тренда $X(t)$ в момент времени t равна среднему арифметическому трех соседних значений ряда: $x_t^* = \frac{x_{t-1} + x_t + x_{t+1}}{3}$. Эта оценка x_t^* называется *сглаженным* значением ряда.

Аналогичным образом можно сгладить все значения ряда, кроме двух крайних значений x_1 и x_n. Разброс сглаженных значений относительно линии истинного тренда должен уменьшаться благодаря усреднению трех случайных компонент ε_{t-1}, ε_t и ε_{t+1}. Данная процедура сглаживания называется методом *простого скользящего среднего*. (В ходе вычислений мы как бы «скользим» вдоль значений ряда, последовательно вычисляя средние x_t^*, затем x^*_{t+1} и т. д.)

Сглаживание временного ряда методом простого скользящего среднего можно проводить с усреднением любого нечетного $2m + 1$ числа значений ряда. Такая процедура будет называться сглаживанием по трем точкам, по пяти точкам, по семи точкам и т. д. А сглаженное по $2m + 1$ точкам значение равно:

$$x_t^* = \frac{x_{t-m} + \ldots + x_t + \ldots + x_{t+m}}{2m+1}. \qquad (7.10)$$

Разброс сглаженных значений ряда x_t^* относительно линии истинного тренда $X(t)$ должен уменьшаться при увеличении числа точек $2m + 1$, по которым производится усреднение. Однако не следует забывать, что внутри интервала $[t - m; t + m]$ истинный тренд должен оставаться практически линейным. Если же сглаживание нелинейного тренда $X(t)$ проводится по избыточному количеству точек, то погрешность (разница между $X(t)$ и x_t^*) наоборот увеличивается.

Число точек $2m + 1$, по которым проводится сглаживание, можно увеличить. Для этого тренд следует аппроксимировать не прямой линией (как в методе простого скользящего среднего), а параболой второй или более высокой степени. В этом случае формула для вычисления x_t^* несколько изменится — каждое усредняемое значение $x_{t-m}, \ldots, x_{t+m}$ следует домножить на постоянный коэффициент, называемый *весом* этого значения. (Отметим, что в формуле (7.10) все веса одинаковы и равны $\frac{1}{2m+1}$.) Такая процедура сглаживания носит название *взвешенного скользящего среднего*.

Приведем (без доказательства) некоторые часто встречающиеся наборы весовых коэффициентов для вычисления взвешенного скользящего среднего.

Сглаживание по пяти точкам, аппроксимация тренда квадратичной или кубической параболой:

$$x_t^* = \frac{-2 \cdot x_{t-2} + 12 \cdot x_{t-1} + 17 \cdot x_t + 12 \cdot x_{t+1} - 2 \cdot x_{t+2}}{35}.$$

Сглаживание по семи точкам, аппроксимация тренда квадратичной или кубической параболой:

$$x_t^* = \frac{-2 \cdot x_{t-3} + 3 \cdot x_{t-2} + 6 \cdot x_{t-1} + 7 \cdot x_t + 6 \cdot x_{t+1} + 3 \cdot x_{t+2} - 2 \cdot x_{t+3}}{21}.$$

Отметим, что весовые коэффициенты симметричны относительно центрального значения, а их сумма всегда равна единице.

Пример. Сгладить временной ряд (7.9) методом простого скользящего среднего с усреднением по трем точкам.

Решение. Первое $x_1 = 0$ и последнее $x_9 = 28$ значения ряда (7.9) не сглаживаются. Простое скользящее среднее для всех остальных точек вычислим по формуле (7.10): $x_2^* = \frac{x_1 + x_2 + x_3}{3} = \frac{0+3+9}{3} = 4$.

$$x_3^* = \frac{x_2 + x_3 + x_4}{3} = \frac{3+9+13}{3} = 8\frac{1}{3},\ x_4^* = \frac{x_3 + x_4 + x_5}{3} = \frac{9+13+16}{3} = 12\frac{2}{3},$$

$$x_5^* = \frac{x_4 + x_5 + x_6}{3} = \frac{13+16+16}{3} = 15,\ x_6^* = \frac{x_5 + x_6 + x_7}{3} = \frac{16+16+24}{3} = 18\frac{2}{3},$$

$$x_7^* = \frac{x_6 + x_7 + x_8}{3} = \frac{16+24+26}{3} = 22,\ x_8^* = \frac{x_7 + x_8 + x_9}{3} = \frac{24+26+28}{3} = 26.$$

Этот ряд представлен на рис. 7.2 ломаной линией, соединяющей сглаженные значения. Исходный ряд (7.9) показан точками.

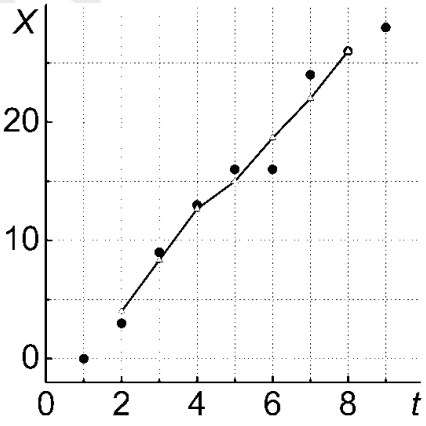

Рис. 7.2. Сглаживание временного ряда (7.9) по методу простого скользящего среднего (усреднение по трем точкам)

Экспоненциальное сглаживание можно применять, когда соседние значения ряда x_{t-1} и x_t разделены настолько малыми промежутками времени, за которые тренд ряда практически не успевает измениться: $X(t-1) \approx \approx X(t)$. Тогда текущим сглаженным значением x_t^* будем считать взвешенное среднее и предыдущего сглаженного значения x^*_{t-1}, и текущего измеренного значения x_t:

$$x_t^* = (1 - C)\cdot x_t + C\cdot x^*_{t-1}, \qquad (7.11)$$

где C — параметр сглаживания. Он должен лежать в интервале $0 < C < 1$, поскольку имеет смысл веса значения x^*_{t-1}.

Формула (7.11) учитывает всю предысторию случайного процесса, все измеренные значения ряда, начиная от первого из них x_1 и до текущего x_t. Действительно, с ее помощью можно выразить предыдущее сглаженное значение $x^*_{t-1} = (1 - C)\cdot x_{t-1} + C\cdot x^*_{t-2}$ и все более ранние сглаженные значения. Подставив их в (7.11), получим: $x_t^* = (1 - C)\cdot x_t + (1 - C)\cdot C\cdot x_{t-1} + \\ + (1 - C)\cdot C^2\cdot x^*_{t-2}$ и т. д. В итоге экспоненциально сглаженное значение x_t^* выразится через более ранние измеренные значения. Если принять, что первое сглаженное значение ряда x_1^* равно первому измеренному значению x_1, то получим:

$$x_t^* = (1 - C)\cdot x_t + (1 - C)\cdot C\cdot x_{t-1} + (1 - C)\cdot C^2\cdot x_{t-2} + \ldots + (1 - C)\cdot C^{t-1}\cdot x_1. \qquad (7.12)$$

Веса измеренных значений $1 - C$, $(1 - C)\cdot C$, $(1 - C)\cdot C^2$, …, $(1 - C)\cdot C^{t-1}$ уменьшаются в геометрической прогрессии по мере их удаления в историю. Это обусловило название данного метода сглаживания. (В геометрической прогрессии убывают значения экспоненты e^{-t} натуральных чисел $t = 1, 2, 3, \ldots$.) Когда параметр C стремится к нулю, сглаживания не происходит, поскольку в этом случае: $x_t^* \to x_t$. Чем больше C, тем больший вес «старых» значений ряда в x_t^*, тем более похожим на «гладкую» горизонтальную линию будет график экспоненциально сглаженного ряда. Если же тренд ряда быстро изменяется, то экспоненциальное сглаживание даст систематическую ошибку — сглаженные значения не будут успевать за трендом. Например, из рис. 7.3 видно, что все значения временного ряда (7.9) лежат выше линии, полученной экспоненциальным сглаживанием с $C = 0{,}5$.

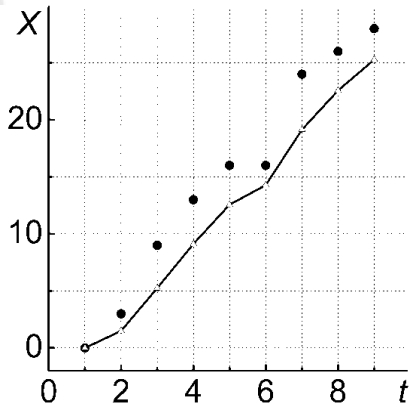

Рис. 7.3. Экспоненциальное сглаживание временного ряда (7.9) с параметром $C = 0{,}5$

7.4. Экстраполяция

Если на данный момент имеются n значений ряда $x_1, x_2, \ldots, x_n$ и случайный процесс продолжает развитие, то нас могут интересовать будущие значения $x_{n+1}, x_{n+2}, \ldots, x_{n+k}$, которые пополнят ряд в моменты времени $n + 1, n + 2, \ldots, n + k$. Прогнозировать будущие, ожидаемые значения случайной величины X можно на основе ее предыдущих, уже известных значений. Напомним, что мы исследуем временные ряды типа (7.5), состоящие из детерминированного тренда $X(t)$ и стационарной случайной компоненты ε_t, математическое ожидание которой равно нулю. Математическое ожидание будущего значения такого ряда x_{n+k} в силу линейности математического ожидания равно:

$$\mu(x_{n+k}) = \mu(X(n+k)) + \mu(\varepsilon_{n+k}).$$

Поскольку детерминированный тренд в момент времени $n + k$ может принять только одно значение $X(n + k)$, то это значение и будет математическим ожиданием тренда: $\mu(X(n + k)) = X(n + k)$. Поэтому $\mu(x_{n+k}) = X(n + k)$, и задача прогнозирования будущих значений временного ряда сводится к предсказанию будущего поведения тренда.

Предсказание значения тренда $X(n + k)$ в будущий момент времени $n + k$ ($k = 1, 2, 3, \ldots$) назовем *экстраполяцией*.

Экстраполировать тренд на будущее можно на основе какой-либо математической модели тренда. Например, описанный в пункте 7.3 метод экспоненциального сглаживания можно применять, когда значение тренда изменяется очень медленно и в последующий момент: $X(n + k) \approx X(n)$. В этом случае экстраполированное значение равно последнему сглаженному: $x^*_{n+1} \approx x^*_n$.

Параметры математической модели тренда можно оценить методом наименьших квадратов (пункт 7.2). Так, если тренд линейный, то его оценка дается уравнением прямой (7.8). Его экстраполяция на момент времени $n + k$ дает: $x^*_{n+k} = \bar{x} + \dfrac{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)\cdot\left(t - \bar{t}\right)}{\sum_{t=1}^{n}\left(t - \bar{t}\right)^2}\cdot\left(n + k - \bar{t}\right)$.

В основе экстраполяции лежит предположение, что характер тренда не изменится по крайней мере до момента $n + k$. Однако чем более дальнюю перспективу мы рассматриваем, тем меньше вероятность сохранения основной тенденции ряда. Поэтому долговременные прогнозы менее надежны, чем кратковременные.

Еще одна неопределенность прогнозов развития тренда обусловлена тем, что модель тренда, используемая для экстраполяции, зависит от того, какую часть имеющегося временного ряда мы используем для построения

этой модели. Так, в пункте 7.2 оценка линейного тренда ряда (7.9) получена с использованием всех девяти значений ряда. Вычислим значение тренда в крайней точке $t = n = 9$, используя эту оценку:

$$X(9) = \frac{107 \cdot t - 85}{30} = \frac{107 \cdot 9 - 85}{30} = \frac{878}{30} = 29\frac{4}{15}.$$

А при сглаживании ряда простым скользящим средним (7.10) линейный тренд оценивают по $2m + 1$ точкам. Если $2m + 1 = 3$, то оценка тренда (7.8) в крайней точке $t = n$ использует $\bar{x} = \frac{x_{n-2} + x_{n-1} + x_n}{3}$ и $\bar{t} = n - 1$:

$$x_n^* = \bar{x} - \frac{\left(x_{n-2} - \bar{x}\right) \cdot (-1) + \left(x_{n-1} - \bar{x}\right) \cdot 0 + \left(x_n - \bar{x}\right) \cdot 1}{(-1)^2 + 0^2 + 1^2} \cdot 1 =$$

$$= \bar{x} - \frac{x_n - x_{n-2}}{2} = \frac{5x_n + 2x_{n-1} - x_{n-2}}{6}.$$

Оценка тренда ряда (7.9) в крайней точке $t = 9$, сделанная по этой формуле, равна: $x_9^* = \frac{5 \cdot 28 + 2 \cdot 26 - 24}{6} = 28.$

ЛИТЕРАТУРА

1. *Основы* высшей математики и математической статистики : учеб. для вузов / И. В. Павлушков [и др.]. М. : ГЭОТАР-Медиа, 2010. 424 с.

2. *Лобоцкая, Н. Л.* Высшая математика : учеб. для вузов / Н. Л. Лобоцкая, Ю. В. Морозов, А. А. Дунаев. Минск : Выш. шк., 1987. 319 с.

3. *Петри, А.* Наглядная медицинская статистика / А. Петри, К. Сэбин ; пер. с англ. В. П. Леонова. 2-е изд., перераб. и доп. М. : ГЭОТАР-Медиа, 2009. 168 с.

4. *Зайцев, В. М.* Прикладная медицинская статистика : учеб. пособие / В. М. Зайцев, В. Г. Лифляндский, В. И. Маринкин. 2-е изд. СПб. : Фолиант, 2006. 432 с.

5. *Медик, В. А.* Статистика в медицине и биологии : руководство : в 2 т. / В. А. Медик ; под ред. Ю. М. Комарова. М. : Медицина, 2000. Т. 1. Теоретическая статистика. 412 с.

6. *Герасимов, А. Н.* Медицинская статистика : учеб. пособие / А. Н. Герасимов. М. : Медицинское информационное агентство, 2007. 480 с.

7. *Кремер, Н. Ш.* Теория вероятностей и математическая статистика / Н. Ш. Кремер. М. : ЮНИТИ-ДАНА, 2000. 543 с.

8. *Вероятность* и математическая статистика : энцикл. / гл. ред. Ю. В. Прохоров. М. : Большая российская энциклопедия, 1999. 912 с.

9. *Капитонов, А. М.* Основы математического анализа и дифференциальных уравнений : учеб.-метод. пособие / А. М. Капитонов. Минск : БГМУ, 2013. 4751 с.

10. http://www.bsmu.by/files/k_fiziki/2012-2/lecture.pdf.

Приложение 1

Биномиальные коэффициенты $C_n^m = \dfrac{n!}{m!(n-m)!}$

(число сочетаний из n по m)

Число положительных исходов m	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2			1	3	6	10	15	21	28	36	45	55	66	78	91	105	120	136	153	171	190
3				1	4	10	20	35	56	84	120	165	220	286	364	455	560	680	816	969	1140
4					1	5	15	35	70	126	210	330	495	715	1001	1365	1820	2380	3060	3876	4845
5						1	6	21	56	126	252	462	792	1287	2002	3003	4368	6188	8568	11628	15504
6							1	7	28	84	210	462	924	1716	3003	5005	8008	12376	18564	27132	38760
7								1	8	36	120	330	792	1716	3432	6435	11440	19448	31824	50388	77520
8									1	9	45	165	495	1287	3003	6435	12870	24310	43758	75582	125970
9										1	10	55	220	715	2002	5005	11440	24310	48620	92378	167960
10											1	11	66	286	1001	3003	8008	19448	43758	92378	184756
11												1	12	78	364	1365	4368	12376	31824	75582	167960
12													1	13	91	455	1820	6188	18564	50388	125970
13														1	14	105	560	2380	8568	27132	77520
14															1	15	120	680	3060	11628	38760
15																1	16	136	816	3876	15504

Число испытаний n

Плотность вероятности стандартного нормального распределения

$$N(u;0;1)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{u^2}{2}}$$

Значение аргумента u равно сумме чисел в первой строке и первом столбце, например на пересечении столбца «1» и строки «0,50» имеем u = 1,50 и N (1,50; 0; 1) = 0,129517596.

	0	1	2	3
0	0,39894228	0,241970725	0,053990967	0,004431848
0,05	0,398443914	0,229882141	0,048792019	0,003809762
0,10	0,396952547	0,217852177	0,043983596	0,003266819
0,15	0,394479331	0,205936269	0,039550042	0,002794258
0,20	0,391042694	0,194186055	0,035474593	0,002384088
0,25	0,386668117	0,182649085	0,031739652	0,002029048
0,30	0,381387815	0,171368592	0,028327038	0,001722569
0,35	0,375240347	0,160383327	0,02521822	0,001458731
0,40	0,36827014	0,149727466	0,02239453	0,001232219
0,45	0,360526962	0,139430566	0,019837354	0,001038281
0,50	0,352065327	**0,129517596**	0,0175283	0,000872683
0,55	0,342943855	0,120009001	0,015449347	0,000731664
0,60	0,333224603	0,110920835	0,013582969	0,000611902
0,65	0,32297236	0,102264925	0,011912244	0,000510465
0,70	0,312253933	0,094049077	0,010420935	0,00042478
0,75	0,301137432	0,086277319	0,009093563	0,000352596
0,80	0,289691553	0,078950158	0,007915452	0,000291947
0,85	0,277984886	0,072064874	0,006872767	0,000241127
0,90	0,26608525	0,065615815	0,005952532	0,000198655
0,95	0,254059056	0,059594706	0,005142641	0,000163256

Значения функции Лапласа $\Phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^u e^{-\frac{u^2}{2}} \cdot du$

Значение аргумента *u* равно сумме чисел в первой строке и первом столбце, например на пересечении столбца «1» и строки «0,15» имеем *u* = 1,15 и Φ(1,15) = 0,374928064.

	0	1	2	3
0	0	0,341344746	0,477249868	0,498650102
0,01	0,003989356	0,343752355	0,477784406	0,498693762
0,02	0,007978314	0,34613577	0,478308306	0,498736127
0,03	0,011966473	0,348494997	0,47882173	0,498777231
0,04	0,015953437	0,35083005	0,479324837	0,498817109
0,05	0,019938806	0,353140944	0,479817785	0,498855793
0,06	0,023922183	0,3554277	0,48030073	0,498893315
0,07	0,02790317	0,357690346	0,480773828	0,498929706
0,08	0,031881372	0,35992891	0,481237234	0,498964997
0,09	0,035856393	0,362143428	0,4816911	0,498999218
0,10	0,039827837	0,364333939	0,482135579	0,499032397
0,11	0,043795313	0,366500487	0,482570822	0,499064563
0,12	0,047758426	0,368643119	0,482996977	0,499095745
0,13	0,051716787	0,370761888	0,483414193	0,499125968
0,14	0,055670005	0,372856849	0,483822617	0,499155261
0,15	0,059617692	**0,374928064**	0,484222393	0,499183648
0,16	0,063559463	0,376975597	0,484613665	0,499211154
0,17	0,067494932	0,378999516	0,484996577	0,499237805
0,18	0,071423716	0,380999893	0,485371269	0,499263625
0,19	0,075345435	0,382976804	0,485737882	0,499288636
0,20	0,079259709	0,38493033	0,486096552	0,499312862
0,21	0,083166163	0,386860554	0,486447419	0,499336325
0,22	0,087064423	0,388767563	0,486790616	0,499359047
0,23	0,090954115	0,390651448	0,487126279	0,499381049
0,24	0,094834872	0,392512303	0,487454539	0,499402352
0,25	0,098706326	0,394350226	0,487775527	0,499422975

Значения функции Лапласа $\Phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^u e^{-\frac{u^2}{2}} \cdot du$

Значение аргумента *и* равно сумме чисел в первой строке и первом столбце, например на пересечении столбца «2» и строки «0,33» имеем *и* = 2,33 и Ф(2,33) = 0,490096924.

	0	1	2	3
0,26	0,102568113	0,396165319	0,488089375	0,499442939
0,27	0,106419873	0,397957685	0,488396208	0,499462263
0,28	0,110261248	0,399727432	0,488696156	0,499480965
0,29	0,114091881	0,401474671	0,488989342	0,499499063
0,3	0,117911422	0,403199515	0,489275890	0,499516576
0,31	0,121719522	0,404902082	0,489555923	0,499533520
0,32	0,125515835	0,406582491	0,489829561	0,499549913
0,33	0,129300019	0,408240864	**0,490096924**	0,499565770
0,34	0,133071736	0,409877328	0,490358130	0,499581108
0,35	0,136830651	0,411492009	0,490613294	0,499595942
0,36	0,140576433	0,413085038	0,490862532	0,499610288
0,37	0,144308755	0,414656549	0,491105957	0,499624159
0,38	0,148027292	0,416206678	0,491343681	0,499637571
0,39	0,151731727	0,417735561	0,491575814	0,499650537
0,40	0,155421742	0,419243341	0,491802464	0,499663071
0,41	0,159097026	0,420730159	0,492023740	0,499675186
0,42	0,162757273	0,422196159	0,492239746	0,499686894
0,43	0,166402179	0,423641490	0,492450589	0,499698209
0,44	0,170031446	0,425066300	0,492656369	0,499709143
0,45	0,173644780	0,426470740	0,492857189	0,499719707
0,46	0,177241890	0,427854963	0,493053149	0,499729912
0,47	0,180822491	0,429219123	0,493244347	0,499739771
0,48	0,184386303	0,430563377	0,493430881	0,499749293
0,49	0,187933051	0,431887882	0,493612845	0,499758490
0,50	0,191462461	0,433192799	0,493790335	0,499767371
0,51	0,194974269	0,434478288	0,493963442	0,499775947

Значения функции Лапласа $\Phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^u e^{-\frac{u^2}{2}} \cdot du$

Значение аргумента *u* равно сумме чисел в первой строке и первом столбце, например на пересечении столбца «2» и строки «0,576» имеем *u* = 2,576 и Φ(2,576) = 0,495002468.

	0	1	2	3
0,52	0,198468212	0,435744512	0,494132258	0,499784227
0,53	0,201944035	0,436991636	0,494296874	0,499792220
0,54	0,205401484	0,438219823	0,494457377	0,499799936
0,55	0,208840313	0,439429242	0,494613854	0,499807384
0,56	0,212260281	0,440620059	0,494766392	0,499814573
0,57	0,215661151	0,441792444	0,494915074	0,499821509
0,576	0,217692409	0,442487098	**0,495002468**	0,499825554
0,58	0,219042691	0,442946567	0,495059984	0,499828203
0,59	0,222404675	0,444082597	0,495201203	0,499834661
0,6	0,225746882	0,445200708	0,495338812	0,499840891
0,61	0,229069096	0,446301072	0,495472889	0,499846901
0,62	0,232371107	0,447383862	0,495603512	0,499852698
0,63	0,235652708	0,448449252	0,495730757	0,499858289
0,64	0,238913700	0,449497417	0,495854699	0,499863681
0,65	0,242153889	0,450528532	0,495975411	0,499868880
0,66	0,245373085	0,451542774	0,496092967	0,499873892
0,67	0,248571105	0,452540318	0,496207438	0,499878725
0,68	0,251747770	0,453521342	0,496318892	0,499883383
0,69	0,254902906	0,454486023	0,496427399	0,499887873
0,70	0,258036348	0,455434537	0,496533026	0,499892200
0,71	0,261147932	0,456367063	0,496635840	0,499896370
0,72	0,264237502	0,457283779	0,496735904	0,499900389
0,73	0,267304908	0,458184862	0,496833284	0,499904260
0,74	0,270350003	0,459070491	0,496928041	0,499907990
0,75	0,273372648	0,459940843	0,497020237	0,499911583
0,76	0,276372708	0,460796097	0,497109932	0,499915043

Значения функции Лапласа $\Phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^u e^{-\frac{u^2}{2}} \cdot du$

Значение аргумента *u* равно сумме чисел в первой строке и первом столбце, например на пересечении столбца «1» и строки «0,96» имеем *u* = 1,96 и Φ(1,96) = 0,475002105.

	0	1	2	3
0,77	0,279350054	0,461636430	0,497197185	0,499918376
0,78	0,282304562	0,462462020	0,497282055	0,499921586
0,79	0,285236116	0,463273044	0,497364598	0,499924676
0,8	0,288144601	0,464069681	0,497444870	0,499927652
0,81	0,291029912	0,464852106	0,497522925	0,499930517
0,82	0,293891946	0,465620498	0,497598818	0,499933274
0,83	0,296730608	0,466375031	0,497672600	0,499935928
0,84	0,299545807	0,467115881	0,497744323	0,499938483
0,85	0,302337457	0,467843225	0,497814039	0,499940941
0,86	0,305105479	0,468557237	0,497881795	0,499943306
0,87	0,307849798	0,469258091	0,497947641	0,499945582
0,88	0,310570345	0,469945961	0,498011624	0,499947772
0,89	0,313267057	0,470621020	0,498073791	0,499949878
0,90	0,315939875	0,471283440	0,498134187	0,499951904
0,91	0,318588745	0,471933393	0,498192856	0,499953852
0,92	0,321213620	0,472571050	0,498249843	0,499955726
0,93	0,323814458	0,473196581	0,498305190	0,499957527
0,94	0,326391220	0,473810155	0,498358939	0,499959259
0,95	0,328943874	0,474411940	0,498411130	0,499960924
0,96	0,331472393	**0,475002105**	0,498461805	0,499962525
0,97	0,333976754	0,475580815	0,498511001	0,499964064
0,98	0,336456941	0,476148236	0,498558758	0,499965542
0,99	0,338912940	0,476704532	0,498605113	0,499966963

Двусторонние критические точки t-распределения Стьюдента $t_{кр}(\alpha; \nu)$
(Коэффициенты Стьюдента $t_{\gamma, \nu}$)

Число степеней свободы $\nu = n - 1$ зависит от объема выборки n.
Уровень значимости α и доверительная вероятность γ связаны соотношением: $\alpha = 1 - \gamma$.
Односторонние критические точки $t_{кр1}(\alpha) = t_{кр}(2\alpha)$, например, $t_{кр1}(0{,}05; 8) = t_{кр}(0{,}1; 8) = 1{,}859548$.

Число степеней свободы $\nu = n - 1$	α	0,2	0,1	0,05	0,02	0,01
	γ	0,8	0,9	0,95	0,98	0,99
1		3,077684	6,313752	12,7062	31,82052	63,65674
2		1,885618	2,919986	4,302653	6,964557	9,924843
3		1,637744	2,353363	3,182446	4,540703	5,840909
4		1,533206	2,131847	2,776445	3,746947	4,604095
5		1,475884	2,015048	2,570582	3,36493	4,032143
6		1,439756	1,94318	2,446912	3,142668	3,707428
7		1,414924	1,894579	2,364624	2,997952	3,499483
8		1,396815	**1,859548**	2,306004	2,896459	3,355387
9		1,383029	1,833113	2,262157	2,821438	3,249836
10		1,372184	1,812461	2,228139	2,763769	3,169273
11		1,36343	1,795885	2,200985	2,718079	3,105807
12		1,356217	1,782288	2,178813	2,680998	3,05454
13		1,350171	1,770933	2,160369	2,650309	3,012276
14		1,34503	1,76131	2,144787	2,624494	2,976843
15		1,340606	1,75305	2,13145	2,60248	2,946713
16		1,336757	1,745884	2,119905	2,583487	2,920782
17		1,333379	1,739607	2,109816	2,566934	2,898231
18		1,330391	1,734064	2,100922	2,55238	2,87844
19		1,327728	1,729133	2,093024	2,539483	2,860935
20		1,325341	1,724718	2,085963	2,527977	2,84534
30		1,310415	1,697261	2,042272	2,457262	2,749996
50		1,298714	1,675905	2,008559	2,403272	2,677793
100		1,290075	1,660234	1,983971	2,364217	2,625891
∞		1,281552	1,644854	1,959964	2,326348	2,575829

Правосторонние критические точки F-распределения Фишера–Снедекора

(Уровень значимости α = 0,05)

Например, если число степеней свободы числителя $\nu_X = 5$ и число степеней свободы знаменателя $\nu_Y = 11$, то правосторонняя критическая точка $F_{кр}(0{,}05; 8; 11) = 3{,}2039$.

Число степеней свободы меньшей дисперсии (знаменателя) $\nu_Y = n_Y - 1$	Число степеней свободы большей дисперсии (числителя) $\nu_X = n_X - 1$								
	1	2	3	4	5	6	7	8	9
1	161,45	199,5	215,71	224,58	230,16	233,99	236,77	238,88	240,54
2	18,513	19	19,164	19,247	19,296	19,33	19,353	19,37	19,385
3	10,128	9,5521	9,2766	9,1172	9,0134	8,9407	8,8867	8,8452	8,812
4	7,7087	6,9443	6,5914	6,3882	6,2561	6,1631	6,0942	6,041	5,9988
5	6,6079	5,7861	5,4095	5,1922	5,0503	4,9503	4,8759	4,8183	4,7725
6	5,9874	5,1433	4,7571	4,5337	4,3874	4,2839	4,2067	4,1468	4,099
7	5,5915	4,7374	4,3468	4,1203	3,9715	3,866	3,7870	3,7257	3,6767
8	5,3177	4,459	4,0662	3,8379	3,6875	3,5806	3,5005	3,438	3,3881
9	5,1174	4,2565	3,8626	3,6331	3,4817	3,3738	3,2928	3,2296	3,1789
10	4,9646	4,1028	3,7083	3,4781	3,3258	3,2172	3,1355	3,0717	3,0204
11	4,8443	3,982	3,5874	3,3567	**3,2039**	3,0946	3,0123	2,948	2,8962
12	4,7472	3,8853	3,4903	3,2592	3,1059	2,9961	2,9134	2,8486	2,7964
13	4,6672	3,8056	3,4105	3,1791	3,0254	2,9153	2,832	2,7669	2,7144
14	4,6001	3,7389	3,3439	3,1123	2,9583	2,8477	2,764	2,6987	2,6458
15	4,5431	3,6823	3,2874	3,0556	2,9013	2,7905	2,7066	2,6408	2,5876
16	4,494	3,6337	3,2389	3,0069	2,8524	2,7413	2,657	2,5911	2,5377
17	4,4513	3,5915	3,1968	2,9647	2,81	2,6987	2,614	2,548	2,4943
18	4,4139	3,5546	3,1599	2,9277	2,7729	2,661	2,5767	2,5102	2,4563
19	4,3808	3,5219	3,1274	2,8951	2,7401	2,6283	2,5435	2,4768	2,4227
20	4,3512	3,4928	3,0984	2,8661	2,7109	2,599	2,514	2,4471	2,3928
21	4,3248	3,4668	3,0725	2,8401	2,6848	2,5727	2,4876	2,4205	2,3661
22	4,301	3,4434	3,0491	2,8167	2,6613	2,5491	2,4638	2,3965	2,3419
23	4,2793	3,4221	3,028	2,7955	2,64	2,5277	2,4422	2,3748	2,3201
24	4,2597	3,4028	3,0088	2,7763	2,6207	2,5082	2,4226	2,3551	2,3002
25	4,2417	3,3852	2,9912	2,7587	2,603	2,4904	2,4047	2,3371	2,2821

Правосторонние критические точки F-распределения Фишера–Снедекора

(Уровень значимости α = 0,1)

Например, если число степеней свободы числителя ν_X = 5 и число степеней свободы знаменателя ν_Y = 11, то правосторонняя критическая точка $F_{кр}(0{,}1; 8; 11) = 2{,}4512$.

Число степеней свободы меньшей дисперсии (знаменателя) $\nu_Y = n_Y - 1$	Число степеней свободы большей дисперсии (числителя) $\nu_X = n_X - 1$								
	1	2	3	4	5	6	7	8	9
1	39,8635	49,5	53,593	55,833	57,24	58,204	58,906	59,439	59,858
2	8,5263	9	9,1618	9,2434	9,2926	9,3255	9,3491	9,3668	9,3805
3	5,5383	5,4624	5,3908	5,3426	5,3092	5,2847	5,2662	5,2517	5,24
4	4,5448	4,3246	4,1909	4,1073	4,0506	4,0098	3,979	3,9549	3,9357
5	4,0604	3,7797	3,6195	3,5202	3,453	3,4045	3,3679	3,3393	3,3163
6	3,776	3,463	3,2888	3,1808	3,1075	3,0546	3,0145	2,983	2,9577
7	3,5894	3,2574	3,0741	2,9605	2,8833	2,8274	2,7849	2,7516	2,7247
8	3,4579	3,1131	2,924	2,8064	2,7265	2,6683	2,6241	2,5894	2,5612
9	3,36	3,0065	2,8129	2,6927	2,6106	2,5509	2,5053	2,4694	2,4403
10	3,285	2,9245	2,7277	2,6053	2,5216	2,4606	2,414	2,3772	2,3473
11	3,225	2,8595	2,6602	2,5362	**2,4512**	2,3891	2,3416	2,304	2,2735
12	3,1766	2,807	2,6055	2,4801	2,394	2,331	2,2828	2,2446	2,2135
13	3,1362	2,7632	2,5603	2,4337	2,3467	2,283	2,2341	2,1954	2,1638
14	3,1022	2,7265	2,5222	2,3947	2,3069	2,2426	2,1931	2,154	2,122
15	3,0732	2,6952	2,4898	2,3614	2,273	2,2081	2,1582	2,1185	2,0862
16	3,0481	2,6682	2,4618	2,3327	2,2438	2,1783	2,128	2,088	2,0553
17	3,0262	2,6446	2,4374	2,3078	2,2183	2,1524	2,1017	2,0613	2,0284
18	3,007	2,624	2,416	2,2858	2,1958	2,1296	2,0785	2,038	2,0047
19	2,99	2,6056	2,397	2,2663	2,176	2,1094	2,058	2,0171	1,9836
20	2,9747	2,5893	2,3801	2,2489	2,1582	2,0913	2,0397	1,9985	1,9649
21	2,961	2,5746	2,3649	2,2333	2,1423	2,0751	2,0233	1,9819	1,948
22	2,9486	2,5613	2,3512	2,2193	2,1279	2,061	2,0084	1,9668	1,9327
23	2,9374	2,5493	2,3387	2,2065	2,1149	2,0472	1,9949	1,9531	1,9189
24	2,9271	2,5383	2,3274	2,1949	2,103	2,0351	1,9826	1,9407	1,9063
25	2,9177	2,5283	2,317	2,1842	2,0922	2,0241	1,9714	1,9293	1,8947

Правосторонние критические точки распределения Пирсона χ^2

Критические точки распределения χ^2 зависят от уровня значимости α (первая строка) и числа степеней свободы $\nu = l - 1 - r$ (первый столбец). Здесь l — количество интервалов группирования, а r — число параметров «теоретического» распределения, оцениваемых с помощью выборки.

Число степеней свободы $\nu = l - 1 - r$	Уровень значимости α		
	0,1	0,05	0,01
1	2,70554	3,84146	6,6349
2	4,60517	5,99146	9,21034
3	6,25139	7,81473	11,3449
4	7,77944	9,48773	13,2767
5	9,23636	11,0705	15,0863
6	10,6446	12,5916	16,8119
7	12,017	14,0671	18,4753
8	13,3616	15,5073	20,0902
9	14,6837	16,919	21,666
10	15,9872	18,307	23,2093
11	17,275	19,6751	24,725
12	18,5493	21,0261	26,217
13	19,8119	22,362	27,6882
14	21,0641	23,6848	29,1412
15	22,3071	24,9958	30,5779
16	23,5418	26,2962	31,9999
17	24,769	27,5871	33,4087
18	25,9894	28,8693	34,8053
19	27,2036	30,1435	36,1909
20	28,412	31,4104	37,5662
21	29,6151	32,6706	38,9322

Правосторонние критические точки распределения Кочрена
(Уровень значимости α = 0,05)

Объем выборок *n*	Количество выборок *k*							
	2	3	4	5	6	7	8	9
2	0,99846	0,96694	0,90646	0,84126	0,78073	0,72698	0,67982	0,63845
3	0,975	0,8709	0,76792	0,68377	0,61615	0,56115	0,51569	0,47749
4	0,93917	0,79774	0,68388	0,59809	0,53212	0,47996	0,4377	0,40274
5	0,9057	0,74566	0,62872	0,54403	0,48035	0,43075	0,39099	0,35838
6	0,87725	0,70699	0,58945	0,50634	0,44472	0,39718	0,35936	0,3285
7	0,85337	0,67704	0,5598	0,47826	0,41841	0,37255	0,33625	0,30675
8	0,83319	0,65305	0,53647	0,45638	0,39802	0,35355	0,31848	0,29008
9	0,81595	0,63331	0,51752	0,43873	0,38167	0,33836	0,30431	0,2768
10	0,80103	0,61672	0,50176	0,42414	0,36818	0,32587	0,29269	0,26594
11	0,78799	0,60253	0,48839	0,41181	0,35684	0,31538	0,28294	0,25684
12	0,77647	0,59022	0,47687	0,40124	0,34712	0,30642	0,27463	0,24909
13	0,76621	0,57942	0,46681	0,39203	0,33869	0,29865	0,26744	0,24238
14	0,75699	0,56984	0,45794	0,38394	0,33129	0,29184	0,26114	0,23652
15	0,74866	0,56127	0,45004	0,37675	0,32473	0,28581	0,25556	0,23133
16	0,74107	0,55355	0,44295	0,37031	0,31885	0,28043	0,25058	0,22671
17	0,73414	0,54655	0,43654	0,36449	0,31356	0,27558	0,24611	0,22255
18	0,72777	0,54016	0,43071	0,35922	0,30876	0,27118	0,24205	0,21879
19	0,72188	0,5343	0,42538	0,3544	0,30439	0,26718	0,23836	0,21536
20	0,71643	0,5289	0,42047	0,34998	0,30038	0,26351	0,23498	0,21223
21	0,71136	0,52391	0,41595	0,3459	0,29668	0,26013	0,23188	0,20935
22	0,70662	0,51927	0,41176	0,34213	0,29326	0,25701	0,229	0,20669
23	0,70219	0,51495	0,40786	0,33862	0,29009	0,25412	0,22634	0,20422
24	0,69804	0,51091	0,40422	0,33536	0,28714	0,25142	0,22387	0,20193
25	0,69412	0,50713	0,40081	0,3323	0,28438	0,24891	0,22155	0,19979

ОГЛАВЛЕНИЕ

Учебное издание

Капитонов Андрей Михайлович

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие

Ответственный за выпуск В. Г. Лещенко
Редактор И. В. Дицко
Компьютерная верстка Н. М. Федорцовой

Подписано в печать 21.03.13. Формат 60×84/16. Бумага писчая «Снегурочка».
Ризография. Гарнитура «Times».
Усл. печ. л. 6,28. Уч.-изд. л. 5,6. Тираж 200 экз. Заказ 656.

Издатель и полиграфическое исполнение:
учреждение образования «Белорусский государственный медицинский университет».
ЛИ № 02330/0494330 от 16.03.2009.
Ул. Ленинградская, 6, 220006, Минск.