1р.60к.

Финансы и статистика

Дж. Полларда

СПРАВОЧНИК ПО ВЫЧИСЛИТЕЛЬНЫМ МЕТОДАМ СТАТИСТИКИ

A HANDBOOK of Numerical and Statistical Techniques

With Examples Mainly From The Life Sciences

J.H.Pollard

Cambridge University Press

Cambridge
London—New York—Melbourne

Дж. Поллард

СПРАВОЧНИК по вычислительным методам статистики

Перевод с английского
В. С. ЗАНАДВОРОВА

Под редакцией
и с предисловием
Е. М. ЧЕТЫРКИНА

Москва
«Финансы и статистика»
1982

ББК 22.172
П51

Поллард Дж.

П51 Справочник по вычислительным методам статистики/ Пер. с англ. В. С. Занадворова; Под ред. и с предисл. Е. М. Четыркина.— М.: Финансы и статистика, 1982.— 344 с., ил.

1 р. 60 к.

В книге дается набор вычислительных процедур численного анализа и статистических приемов оценки гипотез. В ней рассматриваются разного рода ошибки и способы их устранения, методы сглаживания, интерполяция и численное дифференцирование, статистические распределения и моменты, метод наименьших квадратов, анализ регрессий: линейной, криволинейной и нелинейной. Наиболее полно представлена про н практиков, аспирантов и студентов.

Для экономистов и с

П $\frac{0702000000-146}{010(01)-82}$ 37-82

ББК 22.172
517.8

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Статистические методы завоевывают все большее признание как в исследовательской работе, так и в практической деятельности. Это связано с существенными «прорывами» в области статистической методологии, расширением арсенала статистических средств, возможностями реализации соответствующих методик на ЭВМ. Нельзя исключить и то, что произошла, вероятно, переоценка возможностей и ценностей методов количественного анализа экономических явлений, следствием которой явилось известное ослабление внимания к оптимизационным расчетам, некоторый отход от своего рода «оптимизационного романтизма».

Статистический аппарат рассматривается главным образом в учебной литературе, где набор методов, естественно, небогат и стандартен. Описание более тонких и сложных современных статистических методов, как правило, разбросано по многочисленным журнальным статьям и научным публикациям. Все это обусловило необходимость в систематизированном и сжатом изложении методов статистического анализа. Существенную помощь здесь может оказать предлагаемый вниманию советского читателя перевод Справочника по вычислительным методам статистики Дж. Полларда.

Справочник тематически разделен на три части. В первой части рассматриваются некоторые численные методы, необходимость в которых возникает при решении ряда статистических проблем. Помимо чисто математического инструментария Дж. Поллард счел уместным поместить материал, который обычно относят собственно к статистике,— различные приемы сглаживания рядов данных. Здесь же автор приводит малоизвестные формулы, с помощью которых вычисляются концевые сглаженные значения уровней, а также весовые коэффициенты для сглаживания, обеспечивающие оптимальные сглаживающие свойства. Отдельный параграф посвящен специальному методу сглаживания — расчету сплайн-функций. Последние, как известно, существенно улучшают результаты интерполяции, если исходные данные располагаются в ряд, который не может быть достаточно хорошо описан единственной простой функцией. Много внимания уделяется расчетам, базирую-

щимся на конечных разностях. Это вполне оправдано, так как конечные разности широко используются для численных решений дифференциальных уравнений. (Дифференциальные и особенно дифференциально-разностные уравнения стали занимать все более заметное место в моделировании экономических процессов.) Другой важной областью приложений конечных разностей является интерполяция данных.

Во второй части Справочника обсуждаются некоторые основные понятия математической статистики. После сжатого рассмотрения сущности и методов расчета ряда параметров одномерного и двумерного распределений автор переходит к подробной характеристике основных распределений. Помимо обычно встречающихся в учебной литературе распределений здесь приводятся сведения о «менее известных» распределениях, которые для исследователя, применяющего статистические методы, могут оказаться весьма полезными (например, геометрическое и гипергеометрическое распределения, отрицательное биномиальное распределение и т. д.). Специальная глава посвящена характеристике и методам оценки параметров распределений Пирсона.

Наибольшее место в этой части книги занимает важная для исследователя-экспериментатора тема — статистическая проверка гипотез. Приведенный материал отличается широтой охвата методов и постановок задачи и изложен с большим пониманием потребностей практика, занимающегося статистическим анализом. Весьма уместны здесь комментарии, сопровождающие описание отдельных методов. Особенно важны, на наш взгляд, указания на устойчивость того или иного критерия проверки при нарушении исходных предпосылок модели (важный момент, которому в статистической литературе, к сожалению, уделяется недостаточное внимание), а также различного рода предупреждения о возможном неправильном применении метода и т. д.

В этой же части книги четко представлены сведения о методах точечного и интервального оценивания параметров. Опираясь на метод максимального правдоподобия, автор последовательно рассматривает подходы к оцениванию доверительных интервалов для некоторых параметров (и их линейных комбинаций) ряда непрерывных и дискретных распределений.

Третья, заключительная часть Справочника посвящена одной из «классических» тем статистики — регрессионному анализу. Автор приводит материал, демонстрирующий не только технику оценивания параметров соответствующих уравнений, но и статистический анализ полученного уравнения регрессии, что, на наш взгляд, представляется особенно ценным (так как в практике статистический анализ часто заканчивается после получения уравнения).

Читателю следует обратить внимание на замечания и предупреждения, связанные с разработкой регрессий. Однако такие предупреждения сделаны далеко не везде, где, как нам кажется, это необходимо. В частности, автор приводит примеры подбора линейной и криволинейной множественной регрессии к данным наблюдений, состоящим всего из четырех точек, не сопровождая эти расчеты никакими комментариями. Недостаточно полно изложен материал, относящийся к нелинейной по параметрам регрессии [1].

Отличительной особенностью Справочника является большое количество примеров. Собственно, на примерах и происходит обучение. Книга построена по принципу «делай так». Разумеется, для полного понимания ее содержания читателю необходимо обратиться к соответствующей литературе. Заметим, что Дж. Поллард приводит после каждого параграфа список рекомендуемых работ.

Подытоживая сказанное, еще раз отметим, что перевод книги Дж. Полларда представляется своевременным и целесообразным. Издание работы на русском языке отвечает возросшей практической необходимости в более интенсивном и строгом подходе к статистическому анализу экономических явлений. Поскольку Справочник охватывает важные разделы современной статистики, он, без сомнения, окажет большую помощь экономистам — научным работникам и практикам, а также широкому кругу специалистов, имеющих дело с обработкой и анализом статистических данных.

Е. М. ЧЕТЫРКИН

[1] Методы нелинейной регрессии подробно обсуждаются в работе Е. З. Демиденко «Линейная и нелинейная регрессии» (М., Финансы и статистика, 1981).

ПРЕДИСЛОВИЕ

Современные ученые имеют широкий доступ к мощным системам вычислительных машин. Математическое обеспечение подобных систем содержит очень сложные программы, предназначенные для решения статистических и вычислительных задач. Тем не менее многие из повседневных задач, возникающих в процессе научного исследования, могут быть решены (и часто более успешно) с помощью настольных программируемых калькуляторов или мини-компьютеров, которыми теперь оборудована почти каждая лаборатория. По своим вычислительным возможностям эти машины близки к уровню, достигнутому ЭВМ двадцать лет назад; они широко распространяются, так как легки в обращении, дешевы, и к тому же исследователь может «играть» на них со своими данными столько, сколько ему представляется плодотворным. При таком непосредственном контакте с машиной, вероятно, удается достичь значительно больших успехов, чем в тех случаях, когда вычисления производятся в режиме пакетной обработки на мощной ЭВМ. Для настольных машин разработаны комплексы программ их математического обеспечения, однако практика исследований показывает, что обычно для применения этих программ необходимо их модифицировать с учетом конкретных требований. Данная книга призвана помочь исследователю в осуществлении подобной задачи.

Программы, входящие в математическое обеспечение мощных ЭВМ, предназначены для решения стандартных задач. Как правило, они выдают на печать большой объем информации. Для того чтобы использовать эти программы, необходимо ориентироваться в методологии программирования и показателях, которые могут быть представлены в выдаче той или иной программы. Особенно необходимо это в том случае, когда требуется модифицировать или комбинировать готовые программы для решения нестандартных задач. В книге описаны многие из основных численных и статистических методов, программы которых входят в математическое обеспечение.

В настоящее время появляется возможность проводить вычисления с помощью терминалов крупных машин. Эти терминалы обеспечивают работу в диалоговом режиме, позволяя сочетать преимущества мощных вычислительных систем с легкостью обращения, присущей настольным программируемым калькуляторам. Относительно преимуществ работы с такими терминалами можно повторить те же замечания, которые были высказаны по поводу мини-компьютеров и настольных калькуляторов.

Необходимо предостеречь читателя от механического использования стандартных программ. Их рекламные описания нередко могут ввести в заблуждение, а сами программы иногда содержат ошибки. Поэтому мы советуем пользователю проверять на знакомых данных каждую новую для него программу.

Книга написана как практическое руководство для исследователей, знакомых со статистическими методами. Это — не учебник, хотя некоторые методы (например, метод наименьших квадратов) объяснены достаточно подробно. В конце каждой главы приведены упражнения, поясняющие и развивающие отдельные моменты изложения. Применение каждого из описанных методов иллюстрируется примерами. Для решения задачи, возникшей в ходе практического исследования, следует воспользоваться каким-либо из тех методов, которые представляются исследователю целесообразными в этом конкретном случае. Если один метод не подойдет, следует испытать другой. Как правило, среди известных методов всегда можно найти приемлемый. Для решения более трудных вычислительных и статистических задач мы рекомендуем обратиться за консультацией к специалисту по соответствующим методам.

Книга делится на три части. В первой из них рассмотрены численные методы линейной алгебры и анализа, во второй — статистические методы. Третья часть посвящена методу наименьших квадратов, который можно рассматривать и как статистический метод, и как вычислительный прием.

Основная часть этой книги была написана мною в Европе во время научного отпуска, поэтому я был бы рад выразить свою благодарность профессору Г. Фейхтингеру из Вены, профессору Я. Хоему из Копенгагена и профессору П. Уиттлу из Кембриджа за их гостеприимство. Я хотел бы также поблагодарить доктора О. Аалена, доктора Д. Гани, К. Киртона, доктора С. Йохансена и профессора М. Уильямсона за их ценные замечания, а Б. Торн — за выполнение графиков. Эта книга посвящается моей жене, проявившей большое терпение в течение многих вечеров, когда я был занят работой над книгой.

Д Ж. ПОЛЛАРД

Университет Маккуори
Сидней, декабрь 1976 г.

Часть I. ОСНОВНЫЕ ЧИСЛЕННЫЕ МЕТОДЫ

1. ВВЕДЕНИЕ

В данной главе рассмотрены некоторые важнейшие математические результаты, которые зачастую необходимо знать для решения вычислительных и статистических задач, возникающих в ходе научного исследования. Вначале мы рассматриваем разложение в ряд Тейлора некоторых функций одной переменной; затем вводятся понятия частной производной и ряда Тейлора для функций двух или более переменных. В заключительной части главы дается понятие матрицы и обсуждаются основные операции над матрицами — сложение, вычитание, умножение и получение обратной матрицы.

1.1. НЕОБХОДИМЫЙ УРОВЕНЬ МАТЕМАТИЧЕСКОЙ ПОДГОТОВКИ

Для понимания методов, описанных в этой книге, достаточен весьма скромный уровень математических знаний. Тем не менее применение этих методов дает возможность решать сложные вычислительные и статистические задачи. Например, решать нелинейные уравнения с одним или несколькими неизвестными, интегрировать и дифференцировать заданную функцию (в том числе неэлементарную функцию, получаемую в ходе эксперимента), сглаживать исходные данные, подбирать кривую и интерполировать. Некоторые из рассматриваемых методов могут применяться и для решения более сложных задач. Например, метод конечных разностей (см. гл. 6) широко используется для решения дифференциальных уравнений.

Читатель познакомится с математическими понятиями дифференцирования (при определении тангенса угла наклона кривой) и интегрирования (при нахождении площади под кривой). Целесообразно также вспомнить некоторые другие понятия и формулы; они собраны в данной главе.

1.2. РАЗЛОЖЕНИЕ ФУНКЦИИ В РЯД ТЕЙЛОРА

В формуле (1.2.1) значения рассматриваемой функции $f(x)$ и ее производных в заданной точке x используются для получения значения этой функции в близкой точке $x+h$:

$$f(x+h)=f(x)+\frac{h}{1!}f'(x)+\frac{h^2}{2!}f''(x)+\ldots \qquad (1.2.1)$$

Литература: [17, с. 16], [38, с. 407—427], [73, с. 9].

1.3. ЭКСПОНЕНЦИАЛЬНЫЙ РЯД

Следующий ряд для функции e^x сходится к значениям функции для всех значений x:

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots \qquad (1.3.1)$$

Эту формулу можно получить при разложении в ряд Тейлора (1.2.1), вспомнив, что производная функции e^x есть e^x.

Литература: [17, с. 16], [38, с. 428—434, 457—460].

1.4. ЛОГАРИФМИЧЕСКИЙ РЯД

Следующий ряд для функции натурального логарифма от $1+x$ сходится к значениям этой функции при условии, что $|x|<1$:

$$\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\ldots \qquad (1.4.1)$$

Эту формулу можно получить при разложении в ряд Тейлора (1 2 1) вспомнив, что производная функции $\ln(1+x)$ есть $(1+x)^{-1}$.

Литература: [17, с. 16], [38, с. 428—434].

1.5. РАЗЛОЖЕНИЕ ПО ФОРМУЛЕ БИНОМА

Следующий ряд для функции $(1+x)^n$ сходится к значению функции для всех действительных значений n при условии, что $|x|<1$:

$$(1+x)^n=1+\binom{n}{1}x+\binom{n}{2}x^2+\ldots, \qquad (1.5.1)$$

где

$$\left.\begin{aligned}\binom{n}{1}&=\frac{n}{1},\quad \binom{n}{2}=\frac{n(n-1)}{1\times 2},\\ \binom{n}{3}&=\frac{n(n-1)(n-2)}{1\times 2\times 3}\ \text{и т. д.}\end{aligned}\right\} \qquad (1.5.2)$$

Формула может быть получена при разложении в ряд Тейлора (1.2.1).

В случае когда n — положительное целое число, ряд сходится для всех x и содержит лишь $n+1$ ненулевых членов; кроме того, выполняется равенство

$$\binom{n}{r}=\frac{n!}{r!(n-r)!}. \qquad (1.5.3)$$

Численные значения биномиальных коэффициентов (1.5.3) часто располагают в виде *треугольника Паскаля* (см. табл. 1.5.1).

Таблица 1.5.1. **Биномиальные коэффициенты** $\binom{n}{r}$ **для малых значений** n

Каждое число в этой таблице (для $r \geqslant 1$) равно сумме числа, расположенного непосредственно над данным числом, и числа, расположенного над левым, соседним с данным (например, $70 = 35 + 35$).

n	r								
	0	1	2	3	4	5	6	7	8
0	1								
1	1	1							
2	1	2	1						
3	1	3	3	1					
4	1	4	6	4	1				
5	1	5	10	10	5	1			
6	1	6	15	20	15	6	1		
7	1	7	21	35	35	21	7	1	
8	1	8	28	56	70	56	28	8	1

Пример 1.5.1. Воспользуйтесь формулой (1.5.1) для нахождения квадратного корня из 1,03 с точностью до шестого знака после запятой:

$$(1{,}03)^{\frac{1}{2}} = 1 + \frac{\frac{1}{2}}{1!}(0{,}03) + \frac{\frac{1}{2}\left(-\frac{1}{2}\right)}{1\times 2}(0{,}03)^2 +$$

$$+ \frac{\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)}{1\times 2\times 3}(0{,}03)^3 + \ldots = 1 + 0{,}015 - 0{,}000\,112\,5 +$$

$$+ 0{,}000\,001\,6875 - \ldots = 1{,}014\,889 \text{ (с точностью до шестого знака после запятой).}$$

Литература: [38, с. 32—42], [73, с. 8].

1.6. ЧАСТНЫЕ ПРОИЗВОДНЫЕ

Тангенс угла наклона кривой, которая является графиком функции одной переменной $f(x)$, равен скорости возрастания значений этой функции при росте x. График функции двух переменных x и y можно уподобить рельефу участка земной поверхности (в частности, некоторому холму). Значение x можно сопоставить с географической широтой рассматриваемой точки поверхности, значение y — с ее долготой, а высоту — со значением функции $f(x, y)$.

Если бы мы находились у подножия холма и направлялись бы прямо к его вершине, подъем мог бы оказаться довольно трудным из-за значительной крутизны склона. Мы могли бы избрать другой способ подъема и подниматься зигзагом по менее крутым тропинкам. Очевидно, что наклон в заданной точке (x, y) зависит от направления движения.

Возвращаясь к математическому анализу функции двух переменных, выделим два важных направления изменений: первое соответствует возрастанию x при постоянном значении y (постоянная долгота и возрастающая широта), второе — возрастанию y при постоянном x (постоянная широта и возрастающая долгота). Эти два направления всегда образуют прямой угол.

Тангенс угла наклона поверхности в направлении возрастания x при постоянном y называют частной производной функции $f(x, y)$ по x и обозначают символом $\partial f/\partial x$. Ее получают, принимая y за константу и дифференцируя $f(x, y)$ *как* обычную функцию от одной переменной x. Аналогично тангенс угла наклона в направлении возрастания y при постоянном x называют частной производной $f(x, y)$ по y и обозначают $\partial f/\partial y$. Ее получают, принимая x за константу и дифференцируя $f(x, y)$ по y.

На вершине холма угол наклона в любом направлении равен нулю, в том числе и в тех двух направлениях, которые были рассмотрены выше. В точке, где значение функции $f(x, y)$ достигает максимума (или минимума), обе частные производные df/dx и df/dy равны нулю.

Пример 1.6.1. Найти минимальное значение функции

$$f(x, y) = x^2 + y^2 - x + y + xy - 3.$$

Для решения задачи приравняем нулю частные производные по x и y:

$$2x - 1 + y = 0,$$
$$2y + 1 + x = 0.$$

Мы решаем систему из двух уравнений с двумя неизвестными и находим, что минимум достигается в точке $(1, -1)$ и равен -4. Эта стационарная точка является точкой минимума, так как решение системы единственно и $f(x, y) \to \infty$ при $x \to \pm\infty$ или $y \to \pm\infty$.

Литература: [38, с. 435—438], [92, с. 123—126, 160—166].

1.7. ДВУМЕРНЫЙ РЯД ТЕЙЛОРА

Двумерный ряд Тейлора имеет вид

$$f(x + h, y + k) = f(x, y) + \left(h\frac{\partial}{\partial x}f(x, y) + k\frac{\partial}{\partial y}f(x, y)\right) +$$
$$+ \frac{1}{2!}\left(h^2\frac{\partial^2}{\partial x^2}f(x, y) + \right.$$
$$\left. + 2hk\frac{\partial^2}{\partial x\,\partial y}f(x, y) + k^2\frac{\partial^2}{\partial y^2}f(x, y)\right) + \ldots \qquad \textbf{(1.7.1)}$$

Эту формулу можно обобщить на случай трех или более переменных.

Литература: [92, с. 155—157].

1.8. ПОНЯТИЕ МАТРИЦЫ

Матрицей называется прямоугольный массив элементов. Обычно эти элементы являются числами. Например,

$$\begin{pmatrix}-2\\3\end{pmatrix},\quad \begin{pmatrix}1{,}3 & 1{,}2\\1{,}7 & 0{,}6\\0{,}2 & 1{,}3\end{pmatrix},\quad \begin{pmatrix}4{,}3 & 0{,}0\\1{,}9 & 0{,}3\end{pmatrix}.$$

Размерностью матрицы называют число ее строк и столбцов. Так, матрица, состоящая из m строк и n столбцов, называется матрицей размерности $m\times n$. Приведенные выше матрицы имеют размерность 2×1, 3X2 и 2X2 соответственно.

Две матрицы *равны* между собой, если они содержат одно и то же число элементов, расположенных совершенно одинаковым образом, и элементы, находящиеся на одинаковых местах, равны друг другу. Очевидно, что

$$\begin{pmatrix}1 & 2\\3 & 1\end{pmatrix}=\begin{pmatrix}1 & 2\\3 & 1\end{pmatrix} \text{ и } \begin{pmatrix}1 & 2\\3 & 1\end{pmatrix}\neq\begin{pmatrix}1 & 2 & 0\\3 & 1 & 4\end{pmatrix}.$$

Если две матрицы A и B имеют одинаковую размерность, то их *суммой* называют матрицу C, элементы которой получены суммированием соответствующих элементов матриц A и B. Так,

$$A=\begin{pmatrix}1 & 2\\-3 & 4\end{pmatrix},\quad B=\begin{pmatrix}-3 & 2\\4 & 1\end{pmatrix},\quad C=A+B=\begin{pmatrix}-2 & 4\\1 & 5\end{pmatrix}.$$

Аналогично

$$D=A-B=\begin{pmatrix}4 & 0\\-7 & 3\end{pmatrix}.$$

При разной размерности суммы и разности матриц A и B не определяют.

Матрица, все элементы которой равны нулю, называется *нулевой матрицей* и обычно обозначается знаком 0.

При умножении матрицы на заданное число получают матрицу, каждый элемент которой равен соответствующему элементу исходной матрицы, умноженному на это число. Так,

$$-3\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}=\begin{pmatrix}-3 & -6\\-9 & -12\end{pmatrix}=\begin{pmatrix}-1 & -2\\-3 & -4\end{pmatrix}3.$$

Для того чтобы обозначить *произведение двух матриц,* запишем их подряд одну за другой без знака умножения между ними. Например, AB. Произведение AB определяем лишь для тех случаев, когда число столбцов матрицы A равно числу строк матрицы B. Матрица произведения AB имеет столько же строк, сколько A, и столько же столбцов, сколько B. Элемент произведения AB, расположенный в r-й строке и s-м столбце, получают

суммированием попарных произведений элементов r-й строки матрицы A с соответствующими им (по номеру) элементами s-го столбца матрицы B. Так, например, элемент второй строки и первого столбца произведения

$$\text{вторая строка}\rightarrow\begin{pmatrix}2 & 3 & 5\\1 & 9 & 2\\4 & 7 & 6\\14 & 12 & 13\end{pmatrix}\begin{pmatrix}-1 & -5\\12 & 11\\8 & 0\end{pmatrix}$$

$$\uparrow$$

первый столбец

равен $1\times(-1)+9\times 12+2\text{X}8=123$.

В данном случае у первой матрицы размерность 4X3, у второй — 3x2. Матрица произведения имеет размерность 4x2:

$$\begin{pmatrix}74 & 23\\123 & 94\\128 & 57\\234 & 62\end{pmatrix}.$$

Такой способ определения матричного умножения может показаться искусственно усложненным, но он весьма полезен в приложениях, так как мы часто сталкиваемся с необходимостью рассматривать суммы попарных произведений некоторых чисел. В конце данного параграфа приведены соответствующие примеры. Возможна ситуация, когда произведение AB определено, а произведение BA определить нельзя; приведенные выше примеры иллюстрируют эту мысль. Более того, даже если определены оба произведения AB и BA, они в общем случае не равны друг другу.

Квадратная матрица (например, размерности $n\times n$), У которой на *главной диагонали* стоят единицы, а на остальных местах — нули, называется *единичной матрицей* и обычно обозначается символом I*. Например,

$$I=\begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix}.$$

Легко проверить, что во всех случаях, когда произведение определено,

$$AI=IA=A.$$

Говорят, что квадратная матрица A имеет обратную матрицу A^{-1}, если существует такая матрица A^{-1}, что

$$A^{-1}A=AA^{-1}=I.$$

---

* Или E.— *Примеч. ред.*

Квадратная матрица называется *вырожденной,* если для нее нет обратной. Методы нахождения обратной матрицы приведены в параграфе 1.10.

Матрицу размерности 1Хл обычно называют *вектор-строкой, или вектором.* Матрицу размерности $n\times 1$ можно называть *вектор-столбцом,* или *вектором,* К этим векторам применимы описанные выше правила сложения, вычитания и умножения.

Квадратная матрица, у которой все элементы, кроме диагональных, равны нулю, называется *диагональной.* Если все элементы диагонали отличны от нуля, то данная матрица имеет обратную и обратная матрица тоже диагональная. Для получения обратной матрицы надо взять числа, обратные диагональным элементам исходной матрицы. Например,

$$\begin{pmatrix} -2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}.$$

Транспонированную матрицу $\mathbf{A}'$ получают из исходной матрицы А, превращая ее строки в столбцы (и наоборот). Так, например,

$$\text{д} = \begin{pmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{pmatrix} \text{ имеет транспонированную матрицу}$$

$$\mathbf{A}' = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}.$$

Квадратная матрица называется *симметричной,* если $\mathbf{A}' = \mathbf{A}$. Например,

$$\begin{pmatrix} 2 & 5 & 6 \\ 5 & -3 & -7 \\ 6 & -7 & \mathbf{4} \end{pmatrix}.$$

Пример 1.8.1. Одно из важных направлений применения матричной алгебры связано с исследованием систем линейных уравнений. Для того чтобы понять, как это делается, рассмотрим матричное произведение

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 2 \\ 3 & -1 & -2. \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

В результате столбец получается матрица размерности 3x1 или вектор-

$$\begin{pmatrix} * + 2y + 3z \\ 4x \quad + 2z \\ 3x - y - 2z \end{pmatrix}.$$

Если приравнять этот вектор вектор-столбцу

$$\begin{pmatrix} 5 \\ 8 \\ 0 \end{pmatrix},$$

то это будет означать, что мы записали в матричной форме систему из трех уравнений:

$$\left.\begin{aligned} x + 2y + 3z &= 5 \\ 4x \quad + 22 &= \text{i}8 \\ 3x - y - 2z &= \text{b} \end{aligned}\right\}.$$

Обозначим

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 2 \\ 3 & -1 & -2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 5 \\ 8 \\ 0 \end{pmatrix}.$$

А п С известны, а вектор X неизвестен. Для решения системы из трех уравнений необходимо найти вектор **X**, такой, что

$$\mathbf{AX} = \mathbf{C}.$$

Если матрица **A** имеет обратную, то можно умножить $\mathbf{A}^{-1}$ слева на обе части уравнения и получить

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{C}.$$

Это дает нам решение системы из трех линейных уравнений. Следует отметить, что для решения систем линейных уравнений обращение матриц обычно не используется на практике ввиду неэффективности этого метода, однако компактная матричная запись подобных систем удобна для математических выкладок.

Пример 1.8.2. Затраты на новый дом зависят от конструкции и размеров дома, от цены на земельный участок, рабочую силу и материалы. Пусть для постройки дома А необходимы земельный

участок площадью в 10000 единиц, 1000 единиц рабочей силы и 2000 единиц материалов, а для постройки дома B — участок площадью в 7500 единиц, 1500 единиц рабочей силы и 1500 единиц материалов. Эти компоненты затрат можно представить в виде матрицы размерности 3X2:

	Дом A	Дом B	
	10 000	7 500	земельный участок,
	1 000	1 500	рабочая сила,
	2 000	1 500	материалы.

Цены на эти компоненты зависят от местоположения будущего дома. Пусть в данной задаче имеются четыре возможных пункта местоположения (1, 2, 3, 4). Цены каждого вида затрат (выраженные в долларах за соответствующую единицу затрат) представлены в матрице размерности 4x3:

Земельный участок	Рабочая сила	Материалы	
3	4	2	пункт 1,
2	4	3	пункт 2,
2	3	4	пункт 3,
Л	3	5	пункт 4.

Предположим, что некоторый глава семьи, не интересуясь всеми этими деталями, желает просто ознакомиться с совокупными затратами на каждый из двух типов домов в каждом из указанных четырех пунктов. Например, совокупные затраты на дом A в пункте 2 можно подсчитать следующим образом:

$$2X10000 + 4\times1000 + 3X2000 = 30000.$$

Совокупные затраты для обоих домов во всех четырех пунктах могут быть представлены в виде матрицы размерности 4x2; легко видеть, что эта матрица представляет собой произведение двух матриц, приведенных выше. Так,

	Дом A	Дом B						
пункт 1	38 000	31 500		3	4	2		
пункт 2	30 000	25 500		2	4	3		
пункт 3	31 000	25 500		1	2	3	4	
пункт 4	23 000	19 500		1	3	5		

10 000	7 500
1 000	1 500
2 000	1 500

Первая из матриц в правой части задает затраты на единицу материалов, рабочей силы и т. д. При умножении на вторую матрицу (фиксирующую необходимое число таких единиц) получаем матрицу совокупных затрат.

Литература: [17, с. 144—146, 148—150, 152—154], [38, с. 105—135], [41, с. 152], [92, с. 114—120].

1.9. ОПРЕДЕЛИТЕЛИ И АЛГЕБРАИЧЕСКИЕ ДОПОЛНЕНИЯ

Каждой квадратной матрице A можно сопоставить некоторое число, называемое *определителем* или *детерминантом* матрицы. Его обозначают символом det A или символом $|A|$. В случае когда матрица имеет размерность 1x1, определитель просто равен численному значению единственного элемента такой матрицы. Для матрицы размерности 2x2 определитель равен разности двух произведений — произведения элементов, расположенных на главной диагонали, и произведения двух других элементов. Так, например,

$$\begin{vmatrix} 2 & 1 \\ 3 & 4 \end{vmatrix} = 2X4 - 3X1 = 5.$$

Для матриц более высоких размерностей берут некоторую строку или столбец (годится любая строка или столбец). Далее каждый из элементов этой строки или столбца умножают на соответствующее ему число, называемое алгебраическим дополнением, и складывают эти произведения. Полученная сумма является определителем матрицы.

Для того чтобы найти значение *алгебраического дополнения* к элементу матрицы A размерности *nXn,* расположенному в i-й строке и j-м столбце, рассматривают вспомогательную матрицу размерности $(n-1)\times(n-1)$, полученную из исходной матрицы вычеркиванием ее i-й строки и j-го столбца. Если сумма $i+j$ четная, то искомое алгебраическое дополнение равно определителю вспомогательной матрицы. В случае нечетной суммы $i+j$ алгебраическое дополнение получают, умножая этот определитель на -1.

В некоторых случаях рассматривают полную *матрицу алгебраических дополнений к* элементам матрицы A. При этом каждое дополнение записывают на месте, занимаемом соответствующим ему элементом.

Пример 1.9.1. Матрица алгебраических дополнений к элементам матрицы

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \text{ есть } \begin{pmatrix} 4 & -3 \\ -1 & 2 \end{pmatrix}.$$

Правило вычисления определителя матрицы 2X2 было приведено ранее. В действительности это правило является частным случаем более общего правила для матриц *nXn.* Вычислим определитель матрицы A, пользуясь общим правилом. Раскладываем

по первой строке: $\det A = 2\times4 + 1\times(-3) = 5$;
по второй строке: $\det A = 3\times(-1) + 4\times2 = 5$;
по первому столбцу: $\det A = 2X4 + 3\times(-1) = 5$;
по второму столбцу: $\det A = 1\times(-3) + 4\times2 = 5$.

Тот же результат получается при разложении по любой строке и любому столбцу.

Пример 1.9.2. Матрица алгебраических дополнений к элементам матрицы

$$A=\begin{pmatrix}-2 & 5 & 3\\ 4 & 7 & 1\\ 6 & 9 & 8\end{pmatrix}$$

есть

$$\begin{pmatrix}\begin{vmatrix}7&1\\9&8\end{vmatrix} & -\begin{vmatrix}4&1\\6&8\end{vmatrix} & \begin{vmatrix}4&7\\6&9\end{vmatrix}\\ -\begin{vmatrix}5&3\\9&8\end{vmatrix} & \begin{vmatrix}-2&3\\6&8\end{vmatrix} & -\begin{vmatrix}-2&5\\6&9\end{vmatrix}\\ \begin{vmatrix}5&3\\7&1\end{vmatrix} & -\begin{vmatrix}-2&3\\4&1\end{vmatrix} & \begin{vmatrix}-2&5\\4&7\end{vmatrix}\end{pmatrix}=\begin{pmatrix}47 & -26 & -6\\ -13 & -34 & 48\\ -16 & 14 & -34\end{pmatrix}.$$

Для вычисления определителя матрицы A можно воспользоваться любой строкой или любым столбцом. Например, раскладываем

по второму столбцу: $\det A=5\times(-26)+7\times(-34)+9\times 14=-242$;

по третьей строке: $\det A=6\times(-16)+9\times 14+8\times(-34)=-242$.

Читателю предлагается проверить, что результат будет тем же при использовании любой другой строки (или столбца).

Отметим, что для нахождения определителя матрицы нет необходимости вычислять полную матрицу алгебраических дополнений.

Литература: [17, с. 146—148], [38, с. 145—152], [89, с. 347-350], [92, с. 102—109].

1.10. ОБРАЩЕНИЕ МАТРИЦ

Существуют весьма сложные программы обращения матриц, входящие в математическое обеспечение всех ЭВМ; программы обращения разработаны и для многих настольных калькуляторов. Читателю рекомендуется воспользоваться одной из таких программ. В некоторых случаях, однако, оказывается удобным обратить небольшую матрицу вручную. Мы приводим здесь два метода обращения невырожденных квадратных матриц. Большинство обычных программ обращения матриц основано на алгоритмах, близких к первому из описанных нами методов. Второй метод полезен для вычислений вручную при работе с малоразмерными матрицами, однако он малоэффективен (в отношении необходимого числа операций) в случае больших матриц.

В примере 1.8.1 было описано матричное представление системы линейных уравнений. Теперь рассмотрим простую процедуру решения системы из двух линейных уравнений методом исключения одного из неизвестных и построим на ее основе метод обращения матриц 2X2. В приведенной далее записи два исходных уравнения системы записаны под номерами (1) и (2), в каждой строке вслед за очередным уравнением указана соответствующая ему операция процедуры исключения, проделанная для получения этого уравнения; приводится также краткая сводка результатов [1]:

Линейное уравнение	Номер уравнения и операция процедуры исключения	Сводка результатов		
$2x+y=4$	(1)	2	1	4
$3x+4y=11$	(2)	3	4	11
$x+\frac{1}{2}y=2$	$(3)=(1)\div 2$	1	$\frac{1}{2}$	2
$x+\frac{4}{3}y=\frac{11}{3}$	$(4)=(2)\div 3$	1	$\frac{4}{3}$	$\frac{11}{3}$
$x+\frac{1}{2}y=2$	$(5)=(3)$	1	$\frac{1}{2}$	2
$0+\frac{5}{6}y=\frac{5}{3}$	$(6)=(4)-(3)$	0	$\frac{5}{6}$	$\frac{5}{3}$
$x+\frac{1}{2}y=2$	$(7)=(5)$	1	$\frac{1}{2}$	2
$0+y=2$	$(8)=(6)\div\frac{5}{6}$	0	1	2
$x+0=1$	$(9)=(7)-\frac{1}{2}(8)$	1	0	1
$0+y=2$	$(10)=(8)$	0	1	2

Решение системы таково: $x=1$, $y=2$.

Та же последовательность операций может быть и при обращении матрицы

$$A=\begin{pmatrix}2 & 1\\ 3 & 4\end{pmatrix}.$$

Мы используем сводку результатов и подставляем единичную матрицу 2X2 на место столбца чисел, соответствующих правым частям уравнения:

[1] Такой тип сводки результатов рекомендуется при решении системы уравнений вручную.

2	1	1	0	(1)
3	4	0	1	(2)
1	$\frac{1}{2}$	$\frac{1}{2}$	0	(3) = (1) ÷ 2
1	$\frac{4}{6}$	0	$\frac{1}{3}$	(4) = (2) ÷ 3
1	$\frac{1}{2}$	$\frac{1}{2}$	0	(5) = (3)
0	$\frac{5}{6}$	$-\frac{1}{2}$	$\frac{1}{3}$	(6) = (4) − (3)
1	$\frac{1}{2}$	$\frac{1}{2}$	0	(7) = (5)
0	1	$-\frac{3}{5}$	$\frac{2}{5}$	(8) = (6) ÷ $\frac{5}{6}$
1	0	$\frac{4}{5}$	$-\frac{1}{5}$	(9) = (7) − $\frac{1}{2}$ (8)
0	1	$-\frac{3}{5}$	$\frac{2}{5}$	(10) = (8)

Обратная матрица такова:

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{4}{5} & -\frac{1}{5} \\ -\frac{3}{5} & \frac{2}{5} \end{pmatrix}.$$

Легко проверить, что

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

Метод алгебраических дополнений часто применяют при работе с малоразмерными матрицами, но для больших матриц он неэффективен.

1. Напишем матрицу

$$\begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix}.$$

2. Вычислим матрицу алгебраических дополнений (см. параграф 1.9)

$$\begin{pmatrix} 4 & -3 \\ -1 & 2 \end{pmatrix}.$$

3. Вычислим определитель исходной матрицы (умножим элементы одной строки или столбца на соответствующие им элементы матрицы алгебраических дополнений): определитель = $1 \times (-3) + 4 \times 2 = 5$.

4. Разделим элементы матрицы алгебраических дополнений на определитель[2]

$$\begin{pmatrix} \frac{4}{5} & -\frac{3}{5} \\ -\frac{1}{5} & \frac{2}{5} \end{pmatrix}.$$

5. Транспонируем эту матрицу. Получилась обратная матрица

$$\begin{pmatrix} \frac{4}{5} & -\frac{1}{5} \\ -\frac{3}{5} & \frac{2}{5} \end{pmatrix}.$$

Все общие методы обращения матриц включают большое число арифметических операций, поэтому часто возникает проблема необходимой точности, связанная с ошибками округления (см. параграф 2.3). Особенно резко эта проблема проявляется в тех случаях, когда абсолютная величина определителя мала по сравнению с абсолютными величинами некоторых элементов матрицы. В такой ситуации могут оказаться полезными программы арифметических вычислений с двойной точностью (см. параграф 2.3).

Для некоторых видов матриц можно находить обратные довольно простыми способами (например, для диагональных матриц), поэтому обычные программы для обращения таких матриц могут оказаться малоэффективными. В частности, имеются специальные методы для обращения симметричных матриц, такие матрицы часто встречаются *в* статистической практике.

Литература: [17, с. 151, 169—176], [38, с. 136—144], [41, с. 162—163], [89, с. 335—336, 343—347].

1.11. УПРАЖНЕНИЯ

1. Найдите *с* точностью до пяти знаков после запятой значение $(35)^{\frac{1}{5}}$, не обращаясь к таблице логарифмов.

Указание. $(35)^{\frac{1}{5}} = 2\left(1 + \frac{3}{32}\right)^{\frac{1}{5}}$.

2. С помощью формулы (1.5.1) получите разложение функции $(1+x)^{-\frac{3}{2}}$ в ряд по степеням переменной x до члена, содержащего x^3 включительно.

3. Выпишите формулы частных производных по x и по y для функции $f(x, y) = x/y$.

4. Найдите матрицу $\mathbf{Z} = \mathbf{X}(\mathbf{Y} + \mathbf{W})$, где

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 3 & 2 \\ 0 & 0 \\ 1 & 4 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} -2 & 3 \\ 1 & 5 \\ 0 & -2 \end{pmatrix}.$$

5. Найдите матрицу алгебраических дополнений, определитель и обратную матрицу для матрицы A в примере 1.8.1.

6. Воспользуйтесь результатами упражнения 5 для решения системы из трех уравнений, приведенной в примере 1.8.1.

[2] Квадратная матрица с нулевым определителем является *вырожденной*. Для нее не существует обратной матрицы.

2. ПОГРЕШНОСТИ, ОШИБКИ И ОРГАНИЗАЦИЯ ВЫЧИСЛИТЕЛЬНОЙ РАБОТЫ

Результаты вычислений могут отличаться от точных значений соответствующих величин вследствие погрешностей, связанных с отбрасыванием членов ряда, округлением и ошибками в расчетах. В данной главе описаны некоторые простые методы, позволяющие выявить и уменьшить погрешности указанных первых двух типов. Предлагаются также способы проведения процесса вычислений, направленные на устранение ошибок в расчетах.

2.1. ВВЕДЕНИЕ

Результаты вычислений могут отличаться от точных значений величин, соответствующих рассматриваемой математической задаче, по трем основным причинам.

1. Формула, по которой производятся вычисления, возможно, получена путем отбрасывания бесконечного числа членов некоторого ряда и использования конечного числа его первых членов; погрешности этого типа называют *погрешностями отбрасывания членов.*

2. Вычислительные устройства способны фиксировать лишь конечное число десятичных знаков рассматриваемых численных значений, остальные знаки отбрасываются; погрешности этого типа называют *погрешностями округлений.*

3. В ходе расчетов или при записи результатов человек или машина могут ошибаться. Слово *ошибка* использовано нами, чтобы отличить этот источник расхождений, вызванный ошибочными действиями машин или исполнителей, от *«погрешностей»,* вообще говоря, неизбежных, связанных с необходимостью отбрасывать члены ряда или с конечной точностью вычислительных устройств *.

Исследователю следует обеспечить надежность окончательных результатов вычислений, приняв во внимание опасность того, что они могут стать бесполезными из-за ошибок вычислений и слишком больших погрешностей расчетов. В случае многошаговых вычислений рекомендуется проводить промежуточные проверки.

Литература: [17, с. 6—14], [41, с. 1—8], [43, с. 1—8], [62, с. 266—293], [73, с. 2—4], [84, с. 1—8].

2.2. ПОГРЕШНОСТИ, СВЯЗАННЫЕ С ОТБРАСЫВАНИЕМ ЧЛЕНОВ РЯДА

Часто оказывается возможным оценить величину погрешности, порожденной отбрасыванием членов ряда. Подобная оценка позволяет произвести отбрасывание таким образом, чтобы связанные с ним погрешности оставались в приемлемых границах. Так, например, в численных методах интегрирования и дифференцирования обычно используют полиномиальные приближения рассматриваемой функции и в результате получают формулу, содержащую конечное число слагаемых. Эффективность приближения рассматриваемой функции многочленом может быть оценена по разности ординат графиков двух этих функций. Погрешности отбрасывания членов ряда уменьшаются при уменьшении длины отрезка, на котором производится аппроксимация.

Литература: [41, с. 5—8].

* Следует отметить, что слово *ошибка* употребляется и во втором случае; говорят, например, об ошибках отбрасывания членов.— *Примеч. пер.*

2.3. ПОГРЕШНОСТИ ОКРУГЛЕНИЯ

Рассмотрим два числа X и Y, которые представлены в вычислительном устройстве как x и y соответственно. Поскольку значения x и y получены округлением исходных чисел, можно записать:

$$\left.\begin{aligned} x - a < X < x + a \\ y - b < Y < y + b \end{aligned}\right\}, \qquad (2.3.1)$$

где a и b — положительные числа. Очевидно, что

$$x + y - (a + b) < X + Y < x + y + (a + b);$$

$$x - y - (a + b) < X - Y < x - y + (a + b);$$

отсюда можно вывести правило анализа погрешностей: *при сложении (вычитании) двух чисел максимальная возможная погрешность суммы (разности) равна сумме максимальных возможных погрешностей исходных двух чисел.* Это правило учит нас, что следует по возможности избегать в расчетах таких ситуаций, когда вычисляется разность двух больших, близких по значению чисел.

Для того чтобы представить себе величину погрешности при операциях умножения и деления, перепишем неравенства (2.3.1) в виде

$$x(1 - a/x) < X < x(1 + a/x);$$

$$y(1 - b/y) < Y < y(1 + b/y).$$

Предполагается, что максимальные возможные значения относительных ошибок a/x и b/y малы (по сравнению с единицей) и, следовательно, можно пренебречь произведением $(a/x)(b/y)$. Отсюда вытекает, что произведение XY лежит в диапазоне

$$xy\left\{1 \pm \left(\frac{a}{x} + \frac{b}{y}\right)\right\},$$

а частное X/Y — в диапазоне

$$\frac{x}{y}\left\{1 \pm \left(\frac{a}{x} + \frac{b}{y}\right)\right\}.$$

Таким образом, мы вывели правило анализа погрешностей: *при умножении (делении) двух чисел максимальная относительная погрешность произведения (частного) равна сумме максимальных возможных относительных погрешностей.*

Оба сформулированные здесь правила применимы в том случае, когда результат операции не подвергается округлению. При округлении результата необходим дополнительный анализ погрешности. Далее приведен пример такого случая.

В длинной цепи вычислений, включающей в себя много чисел, следует ожидать, что реальная погрешность окончательного результата будет значительно меньше, чем та верхняя граница погрешности, которая задается приведенными правилами, так как некоторые погрешности будут положительны, а другие — отрицательны. Вероятностный подход к погрешностям округления представлен, например, в книге Ральстона [84, с. 8—11].

Большинство современных настольных калькуляторов обеспечивает точность вычислений до двенадцати и более разрядов. Современные ЭВМ обычно обеспечивают девять разрядов (в десятичной системе счисления), и вспомогательные подпрограммы этих ЭВМ дают возможность осуществлять арифметические операции с двойной точностью (что обеспечивает восемнадцать разрядов). При такой высокой точности операций погрешности округления, как правило, мало влияют на общую точность вычислений. Однако в некоторых случаях их роль оказывается весьма существенной, особенно в вычислениях, связанных с большими матрицами, близкими к вырожденным. В сомнительных случаях можно сопоставить результаты вычислений, проделанных дважды, с одинарной и двойной точностью арифметических операций. Обычно величина расхождения этих результатов является хорошим индикатором значимости фактора округления.

Пример 2.3.1. Калькулятор, работающий в режиме вычислений с плавающей запятой, имеет сумматор емкостью в пять десятичных разрядов. После выполнения любой операции результат округляется до ближайшего пятиразрядного числа. Значение выражения $(1{,}05)^{32}$ должно быть оценено последовательным возведением в квадрат содержимого сумматора. Определите ответ, который даст машина, и границы диапазона, заключающего истинное значение выражений.

Задачу можно решать следующим образом:

$$(1{,}05)^2 = 1{,}1025 \text{ (точно)}$$

$$(1{,}05)^4 = (1{,}1025)^2 = 1{,}2155 \pm 0{,}000\,05 \qquad (\pm 0{,}004\,\%)$$

$$(1{,}05)^8 = (1{,}2155 \pm 0{,}004\,\%)^2 = (1{,}2155)^2 \pm 0{,}008\,\% =$$
$$= (1{,}4774 \pm 0{,}000\,05) \pm 0{,}008\,\% = 1{,}4774 \pm 0{,}000\,17 \qquad (\pm 0{,}012\,\%)$$

$$(1{,}05)^{16} = (1{,}4774 \pm 0{,}012\,\%)^2 = (1{,}4774)^2 \pm 0{,}024\,\% =$$
$$= (2{,}1827 \pm 0{,}000\,05) \pm 0{,}024\,\% = 2{,}1827 \pm 0{,}000\,57 \qquad (\pm 0{,}026\,\%)$$

$$(1{,}05)^{32} = (2{,}1827 \pm 0{,}026\,\%)^2 = (2{,}1827)^2 \pm 0{,}052\,\% =$$
$$= (4{,}7642 \pm 0{,}000\,05) \pm 0{,}052\,\% = 4{,}7642 \pm 0{,}00253 \qquad (\pm 0{,}053\,\%).$$

Машина даст ответ 4,7642, истинное значение должно лежать в диапазоне $4{,}7642 \pm 0{,}0025$. Заметим, что $(1{,}05)^{32} = 4{,}7649$ (с точностью до четырех знаков после запятой).

Литература: [41, с. 5—8], [43, с. 14—17], [84, с. 8—10, 12—17].

2.4. ОШИБКИ И ОРГАНИЗАЦИЯ ВЫЧИСЛИТЕЛЬНОЙ РАБОТЫ

Исследователь, несомненно, постарается самым тщательным образом зафиксировать результаты своих экспериментов. Он должен распространить подобную аккуратность на выполнение вычислительной работы. Не следует производить вычисления на случайных клочках бумаги, запись вычислений надо вести в определенном порядке и таким образом, чтобы можно было восстановить, как получены промежуточные и окончательные результаты. Такой подход позволит избежать значительного количества ошибок, а также выявить и исправить те немногие ошибки, которые все-таки будут сделаны.

Часто возникают ошибки при переписывании чисел с одной страницы на другую. Поэтому предпочтительнее вести записи на несброшюрованных листах, а не в журнале, так как такие листы можно положить вплотную друг к другу в момент переписывания. Во многих случаях целесообразно фотокопировать колонки чисел и приклеивать копию на очередной рабочий лист с записями вычислений.

При переписывании чисел весьма типичны следующие две ошибки. Первая связана с изменением взаимного расположения двух соседних цифр; например, число 358 764 может быть записано вместо 385 764. Вторая встречается при записи чисел, в которых идут подряд две цифры и удваивается цифра, соседняя к двойной; например, пишут 48 835 вместо 48 335. Метод обнаружения и исправления чисел в таблице дан в параграфе 6.3.

Почти бесполезны проверки вычислений, если их производит тот же человек при той же схеме расчетов. Как правило, люди склонны многократно повторять одни и те же ошибки. Во всех случаях, когда это возможно, для проверки следует прибегать к альтернативной схеме расчетов. Если все же приходится использовать ту же схему, то проверочные вычисления должны быть поручены другому лицу.

Литература: [41, с. 8—9].

2.5. УПРАЖНЕНИЯ

1. Предполагается, что следующие вычисления производятся на пятиразрядном калькуляторе, описанном в примере 2.3.1:

$$\frac{100{,}23}{300{,}59} - \frac{200{,}47}{601{,}19}.$$

Определите, какой ответ даст машина, и укажите, в каких границах должно лежать истинное значение выражения. Предположите, что исходные числа являются точными.

2. Решите вычислительную задачу из упражнения 1 в предположении, что калькулятор обеспечивает не пять, а четыре значащие цифры.

3. Предложите способ вычислений, реализуемый на пятиразрядном калькуляторе и обеспечивающий решение вычислительной задачи из упражнения 1 с точностью до трех или более значащих цифр.

4. На описанном выше пятиразрядном калькуляторе необходимо вычислить значение выражения $e^{-0,5}$ по формуле (1.3.1). Определите, какой ответ даст машина, и укажите, в каких границах должно лежать истинное значение. Сравните эти результаты с истинным значением.

3. ВЕЩЕСТВЕННЫЕ КОРНИ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

В этой главе представлены различные методы нахождения вещественных корней нелинейных уравнений. У каждого из них есть свои достоинства и недостатки. Метод ложного положения (см. параграф 3.2) — единственный из них, который всегда сходится. Метод Ньютона—Рафсона, описанный в параграфе 3.6 и предназначенный для решения системы из двух нелинейных уравнений, может быть обобщен на случай трех и более переменных.

3.1. ВВЕДЕНИЕ

В ходе научной работы часто приходится иметь дело с нелинейными уравнениями. Для решения подобных уравнений могут оказаться полезными графические построения, но часто требования к порядку точности таковы, что графическое решение рассматривается лишь как первое приближение. Могут также применяться методы проб и ошибок. Обычно наиболее удобны *итеративные методы*. Так называют методы, использующие оценку значения функции в точке x_{n-1} для того, чтобы получить значение x_n (n-го приближения к корню уравнения).

Итеративные процедуры, описанные в данной главе, обычно вполне удовлетворительны, но читатель должен иметь в виду, что в некоторых случаях они могут оказаться непригодными, например, при наличии кратных корней или корней, близких друг к другу. Исследователю рекомендуется при решении конкретной задачи применить тот метод, который на его взгляд приемлем. Если метод окажется непригодным, следует испытать другой. Если подходящий метод не будет найден, мы советуем обратиться за помощью к специалисту по численному анализу.

Литература: [17, с. 19—70], [28, с. 17—26], [41, с. 190—197], [43, с. 443], [73, с. 169—223], [92, с. 95—96].

3.2. МЕТОД ЛОЖНОГО ПОЛОЖЕНИЯ *

Предположим, что необходимо решить нелинейное уравнение $f(x)=0$ и мы установили, что значение $f(a)$ положительно, а $f(b)$ отрицательно. Искомый корень лежит в промежутке между a и b. Поэтому проверим некоторое значение из этого промежутка, но какое значение? Если $f(b)$ ближе к нулю, чем $f(a)$, то было бы логично взять значение, которое ближе к b, чем к a, и наоборот.

* В советской литературе часто используется название «метод вилки» (см., например: Бахвалов Н. С. Численные методы. М., Наука, 1975, с. 419).— *Примеч. пер.*

Один из подходов предлагает выбрать новое значение c при линейной интерполяции:

$$c = \frac{f(a)\,b - f(b)\,a}{f(a) - f(b)}. \qquad (3.2.1)$$

Далее повторяем эту процедуру, используя c как одну из рассмотренных выше двух точек, а в качестве второй точки берется та из точек a и b, которая в паре с c заключает отрезок, содержащий корень.

В отличие от других методов, описанных в данной главе, этот метод всегда сходится, но при этом обычно требуется большее число итераций, чем для остальных методов. Вблизи окончательного решения могут возникнуть проблемы, связанные с погрешностями округления, так как приходится рассматривать разности двух близких чисел.

Пример 3.2.1 [1]. Решим квадратное уравнение

$$f(x) = x^2 - 2 = 0.$$

Мы начинаем проверкой значений:

$$a = 1{,}5; \quad f(a) = 0{,}25;$$
$$b = 1{,}4; \quad f(b) = -0{,}04.$$

Новая точка такова:

$$c = \{0{,}25 \times 1{,}4 - (-0{,}04) \times 1{,}5\}/\{0{,}25 - (-0{,}04)\} = 1{,}4138.$$

Теперь повторяем процедуру, используя значения

$$a = 1{,}5; \quad f(a) = 0{,}25;$$
$$b = 1{,}4138; f(b) = -0{,}0012.$$

Новой точкой будет

$$c = \{0{,}25 \times 1{,}4138 - (-0{,}0012) \times 1{,}5\}/\{0{,}25 - (-0{,}0012)\} = 1{,}4142.$$

При дальнейшем повторении процедуры получен тот же ответ (с точностью до четырех знаков после запятой). Решение найдено.

Литература: [43, с. 446], [100, с. 92; русский перевод с. 89—91].

3.3. МЕТОД НЬЮТОНА—РАФСОНА

Предположим, что необходимо решить нелинейное уравнение $f(x)=0$. Обозначим неизвестное решение через x_∞. Начинаем процесс решения с исходного значения x_0, которое может быть най-

[1] Этот пример был выбран из-за простоты. Математическое обеспечение современных ЭВМ и калькуляторов включает программы, автоматизирующие извлечение квадратного корня.

дено методом проб и ошибок. Если мы сумеем найти поправку h_0, такую, что

$$x_0 + h_0 = x_\infty,$$

то получим решение нелинейного уравнения.

Исходя из разложения Тейлора (1.2.1), можно записать:

$$0 = f(x_\infty) = f(x_0 + h_0) \approx f(x_0) + f'(x_0),$$

откуда получаем

$$h_0 \approx -f(x_0)/f'(x_0),$$

и следующее приближение к корню задается формулой

$$x_1 = x_0 - f(x_0)/f'(x_0).$$

Теперь повторяем эту процедуру, взяв x_1 вместо x_0, чтобы получить приближение x_2, которое еще лучше, чем *XL* Процесс продолжается до тех пор, пока не будет получен корень уравнения с заданной степенью точности.

Этот итеративный процесс можно представить уравнением

$$x_{n+1} = x_n - f(x_n)/f'(x_n). \qquad (3.3.1)$$

Пример 3.3.1 [2]. Решим квадратное уравнение

$$f(x) = x^2 - 2 = 0.$$

Заметим, что $f'(x) = 2x$. Теперь проверим $x_0 = 1{,}5$ как исходное значение. Далее,

$$x_1 = 1{,}5 - f(1{,}5)/f'(1{,}5) = 1{,}5 - (0{,}25)/(3{,}0) =$$
$$= 1{,}4167 \text{ (до четырех знаков после запятой).}$$

Второе приближение таково:

$$x_2 = 1{,}4167 - f(1{,}4167)/f'(1{,}4167) = 1{,}4167 - (0{,}00704)/(2{,}8334) =$$
$$= 1{,}4142 \text{ (до четырех знаков после запятой).}$$

При дальнейшем повторении процедуры получаем тот же ответ; таким образом, корень уравнения (с точностью до четырех знаков после запятой) равен 1,4142.

Пример 3.3.2. Численность популяции ящериц можно оценить по данным об их поимках (с учетом повторных поимок). Обозначим число ящериц, пойманных n раз, величиной $f_n (n = 1, 2, \ldots)$. Как следует из работы Крэйга [18], оценочное значение численности популяции $\hat{P}$ может быть вычислено по формуле

$$\ln(\hat{P}) - \ln(\hat{P} - \sum f_n) = \left(\sum n f_n\right)/\hat{P}. \qquad (3.3.2)$$

[2] См. сноску 1.

Студентами университета Маккуори, изучавшими экологию, в мае 1974 г. были получены следующие данные о популяции ящериц на озере Смит:

n	1	2	3	$\geqslant 4$
f_n	91	29	1	0

Применим метод Ньютона—Рафсона для нахождения $\hat{P}$. Уравнение (3.3.2) можно записать в виде

$$f(\hat{P}) = 0,$$

где

$$f(\hat{P}) = \ln(\hat{P}) - \ln(\hat{P} - \sum f_n) - \left(\sum n f_n\right)/\hat{P}$$

и

$$f'(\hat{P}) = \frac{1}{\hat{P}} - \frac{1}{\hat{P} - \sum f_n} + \left(\sum n f_n\right)/\hat{P}^2.$$

Заметим, что

$$\sum f_n = 91 + 29 + 1 = 121,$$
$$\sum n f_n = 1 \times 91 + 2 \times 29 + 3 \times 1 = 152.$$

В качестве начального значения примем $\hat{P}_0 = 300$. Улучшенная оценка задается формулой

$$\hat{P}_1 = 300 - (5{,}703\,78 - 5{,}187\,39 - 0{,}506\,67) \div$$
$$\div (0{,}003\,33 - 0{,}005\,59 + 0{,}001\,69) = 317.$$

Второе приближение задается формулой

$$\hat{P}_2 = 317 - (5{,}758\,90 - 5{,}278\,11 - 0{,}479\,50) \div$$
$$\div (0{,}003\,15 - 0{,}005\,10 + 0{,}001\,51) = 320,$$

третье приближение — формулой

$$\hat{P}_3 = 320 - (5{,}768\,32 - 5{,}293\,30 - 0{,}475\,00) \div$$
$$\div (0{,}003\,13 - 0{,}005\,03 + 0{,}001\,48) = 320.$$

Значение 320 представляет собой оценку численности популяции ящериц.

Литература: [17, с. 30—39], [28, с. 19—26], [41, с. 193], [42, с. 77—90], [43, с. 447], [73, с. 171—176], [84, *с.* 332], [92, с. 97—102].

3.4. МЕТОД СЕКУЩЕЙ

При применении метода Ньютона—Рафсона необходимо оценивать значение производной для каждого из последовательных приближений. Иногда это трудно и часто утомительно. В методе секущей производная заменена отношением

$$[f(x_n) - f(x_{n-1})]/(x_n - x_{n-1})$$

и используется следующее рекуррентное соотношение:

$$x_{n+1} = x_n - (x_n - x_{n-1}) f(x_n)/(f(x_n) - f(x_{n-1})). \qquad (3.4.1)$$

Это уравнение идентично уравнению (3.2.1), в котором показаны a, b и c вместо x_n, x_{n-1} и x_{n+1} соответственно. Но эти два метода не тождественны друг другу. Метод ложного положения требует, чтобы значения $f(x_n)$ и $f(x_{n-1})$ имели противоположные знаки; в методе секущей $f(x_n)$ и $f(x_{n-1})$ иногда имеют один и тот же знак.

Пример 3.4.1 [3]. Решим квадратное уравнение

$$f(x) = x^2 - 2 = 0.$$

Методом проб и ошибок можно установить, что значения $x_0 = 1{,}5$ и $x_1 = 1{,}4$ лежат вблизи корня уравнения. С помощью (3.4.1) имеем

$$x_2 = 1{,}4 - (1{,}4 - 1{,}5)(-0{,}04)/\{(-0{,}04) - 0{,}25\} = 1{,}4138;$$

$$x_3 = 1{,}4138 - (1{,}4138 - 1{,}4)(-0{,}0012) \div \{(-0{,}0012) - (-0{,}04)\} = 1{,}4142.$$

При повторении этой процедуры получаем тот же ответ. Таким образом, корень уравнения (с точностью до четырех знаков после запятой) равен 1,4142.

Литература: [17, с. 39—40], [84, с. 323—328].

3.5. ПРОСТЫЕ ИТЕРАТИВНЫЕ МЕТОДЫ

Иногда простое преобразование нелинейного уравнения приводит к вполне приемлемому рекуррентному соотношению. Рекуррентные соотношения в параграфах 14.9—14.12, предназначенные для оценки параметров цензурированных и усеченных распределений * (нормального и пуассоновского), были получены подобным образом. Следующий пример окажется полезным при исследовании свойств некоторого распределения, он был впервые приведен в работе А. Лотки в 1931 г.

Пример 3.5.1. Найдем корень полиномиального уравнения, который лежит между нулем и единицей:

$$x = 0{,}4982 + 0{,}2103x + 0{,}1270x^2 + 0{,}0730x^3 + 0{,}0418x^4 + 0{,}0241x^5 + 0{,}0132x^6 + 0{,}0069x^7 + 0{,}0035x^8 + 0{,}0015x^9 + 0{,}0005x^{10}. \qquad (3.5.1$$

Сумма коэффициентов правой части уравнения равна единице; следовательно, $x = 1$ является корнем данного уравнения. Если в выражении (3.5.1) мы заменим x на $(1 - y)$, то получим полиномиальное уравнение относительно y, не содержащее свободного члена. Не интересующий нас корень $y = 0$ можно устранить, разделив обе части уравнения на y. Полученное уравнение относительно y теперь преобразуем к виду

$$y = 0{,}107\,00 + 1{,}009\,96y^2 - 0{,}841\,80y^3 + 0{,}560\,21y^4 - 0{,}288\,15y^5 + 0{,}109\,87y^6 - 0{,}029\,15y^7 + 0{,}004\,80y^8 - 0{,}000\,37y^9. \qquad (3.5.2)$$

Известно, что искомый корень y лежит ближе к нулю, чем к единице. Первое приближение к корню можно найти, отбросив в уравнении (3.5.2) все члены, содержащие y, в степени выше двух и решив получившееся квадратное уравнение

$$1{,}009\,96y^2 - y + 0{,}107\,00 = 0. \qquad (3.5.3)$$

Получаем первое приближение $y = 0{,}122$. Подставляя это значение в правую часть уравнения (3.5.2), получаем второе приближение $y = 0{,}12061$. Вновь с помощью соотношения (3.5.2) получаем третье и четвертое приближения: значения $y = 0{,}120\,34$ и $y = 0{,}12027$ соответственно. Отсюда заключаем, что искомым корнем уравнения (3.5.1) является значение $x = 0{,}8797$ (с точностью до четырех знаков после запятой).

Литература: [60], [61], [73, с. 177—178], [83, с. 102—103].

3.6. ДВУМЕРНЫЙ МЕТОД НЬЮТОНА—РАФСОНА

Предположим, что необходимо решить систему из двух нелинейных уравнений

$$f_1(x, y) = 0 \quad \text{и} \quad f_2(x, y) = 0, \qquad (3.6.1)$$

неизвестное решение обозначим через (x_∞, y_∞). Начинаем процесс решения с исходной точки (x_0, y_0), которую можно отыскать методом проб и ошибок. Если нам удастся найти поправки h_0 и k_0, такие, что

$$x_0 + h_0 = x_\infty \quad \text{и} \quad y_0 + k_0 = y_\infty,$$

то получим решение системы нелинейных уравнений. В общем случае на n-й итерации найдена точка (x_n, y_n), и мы ищем поправки h_n и k_n, такие, что

$$x_n + h_n = x_\infty \quad \text{и} \quad y_n + k_n = y_\infty. \qquad (3.6.2)$$

Используя двумерный ряд Тейлора (1.7.1) и приравнивая к нулю разложения функций f_1 и f_2 в точке $(x_n + h_n, y_n + k_n)$, получаем

$$\left.\begin{aligned} \left(\frac{\partial}{\partial x} f_1(x_n, y_n)\right) h_n + \left(\frac{\partial}{\partial y} f_1(x_n, y_n)\right) k_n &= -f_1(x_n, y_n) \\ \left(\frac{\partial}{\partial x} f_2(x_n, y_n)\right) h_n + \left(\frac{\partial}{\partial y} f_2(x_n, y_n)\right) k_n &= -f_2(x_n, y_n) \end{aligned}\right\}. \qquad (3.6.3)$$

[3] См. сноску 1.

* См. параграф 14.8— *Примеч. ред.*

Далее ищем решение полученной системы из двух линейных уравнений с неизвестными h_n и k_n и получаем улучшенный вариант решения системы нелинейных уравнений (3.6.1):

$$\left.\begin{aligned} x_{n+1} &= x_n + h_n \\ y_{n+1} &= y_n + k_n \end{aligned}\right\}. \qquad (3.6.4)$$

Этот процесс продолжается до тех пор, пока не будет достигнута желаемая степень точности.

Аналогично производится обобщение метода на случай трех и более переменных.

Пример 3.6.1. Решим систему из двух нелинейных уравнений:

$$x^2 + (y - 4)^2 - 9 = 0,$$
$$(x - 4)^2 + y^2 - 9 = 0,$$

используя начальные значения $x_0 = 2{,}4$ и $y_0 - 2{,}6$.

Заметим, что

$$\frac{\partial f_1}{\partial x} = 2x, \quad \frac{\partial f_2}{\partial x} = 2(x - 4);$$

$$\frac{\partial f_1}{\partial y} = 2(y - 4), \quad \frac{\partial f_2}{\partial y} = 2y.$$

Таким образом, на первой итерации необходимо решить систему линейных уравнений

$$4{,}8h_0 - 2{,}8k_0 = 1{,}28;$$
$$-3{,}2h_0 + 5{,}2k_0 = -0{,}32.$$

Находим, что $h_0 = 0{,}36$ и $k_0 = 0{,}16$, таким образом, улучшенное приближение задается значениями $x_1 = 2{,}76$ и $y_1 = 2{,}76$.

При повторении этой операции получены результаты:

n	x_n	y_n
2	2,7089	2,7089
3	2,7071	2,7071
4	2,7071	2,7071

Следовательно, решением являются значения $x = y = 2{,}7071$ (этот результат можно проверить с помощью графиков).

Литература: [17, с. 45—46], [42, с. 105—108], [43, с. 451], [73, с. 200—210], [100, с. 90—92; русский перевод с. 87—89].

3.7. УПРАЖНЕНИЯ

1. Примените метод ложного положения для решения уравнения (с точностью до четырех знаков после запятой)

$$e^x + \ln x = 3{,}4.$$

Решение лежит в промежутке между значениями $x = 1{,}1$ и $x = 1{,}2$.

2. Решите уравнение из упражнения 1 с помощью метода Ньютона—Рафсона.

3. Примените двумерный метод Ньютона—Рафсона для решения системы нелинейных уравнений:

$$\frac{1}{2}x^2 + \ln y = 1{,}3;$$

$$\frac{1}{2}y^2 + \ln x = 0{,}825.$$

Ищется решение в окрестности точки (2; 0,6).

4. ПРОСТЫЕ МЕТОДЫ СГЛАЖИВАНИЯ ИСХОДНЫХ ДАННЫХ

В данной главе описаны два простых метода, предназначенные для сглаживания исходных данных: метод скользящего среднего и сглайн-метод.

4.1. ВВЕДЕНИЕ

Известно очень много методов, предназначенных для сглаживания и выравнивания исходных данных, а также для подбора кривой, соответствующей этим данным. По-видимому, чаще всего применяют графический метод, особенно при подборе прямой линии; многие нелинейные задачи могут быть сведены к линейным путем соответствующего преобразования показателей или использования логарифмической шкалы. Некоторые методы являются узкоспециализированными. В данной главе описаны два простых метода. Важнейший из методов — метод наименьших квадратов, которому посвящена вся третья часть данной книги.

1.2. ФОРМУЛА СГЛАЖИВАНИЯ СКОЛЬЗЯЩИМ СРЕДНИМ

На протяжении данного параграфа удобно представить любое наблюдаемое значение y_i как сумму двух компонент: истинного (или основного) значения *f(Xi)* и накладывающейся на него статистической ошибки e_i. Таким образом,

$$y_i = f(x_i) + e_i.$$

При сглаживании ряда данных наша цель заключается в том, чтобы уменьшить, насколько это возможно, ошибки $\{e_i\}$.

К статистике, экономических, биологических и прочих исследованиях используют *скользящие* (или *подвижные*) *средние*. Применим этот метод к данным наблюдений, приведенных в табл. 1.2.1. Воспользуемся формулой скользящего среднего

$$Y_i = \frac{1}{5}(y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2}).$$

Символом Y_i обозначено сглаженное значение в точке *i*. Результаты процедуры приведены на графике (рис. 4.2.1), демонстрирующем достигнутое сглаживание.

Таблица 4.2.1. **Сглаживающий эффект скользящего среднего**

x	Наблюдаемое значение y	Сглаженное значение Y	x	Наблюдаемое значение y	Сглаженное значение Y
35	95	—	53	1274	1425
36	638	—	54	1579	1500
37	191	324	55	1873	1642
38	419	381	56	1482	1789
39	278	331	57	2002	2039
40	381	425	58	2011	2009
41	384.	437	59	2827	2301
42	665	584	60	1722	2575
43	477	727	61	2942	2720
44	1015	822	62	3374	2701
45	1093	845	63	2735	3221
45	860	922	64	2731	3618
47	779	909	65	4323	3766
48	862	864	66	4926	4039
49	951	913	67	4114	4659
50	866	1016	68	4099	5015
51	1109	1098	69	5835	—
52	1291	1224	70	6103	—

При применении данного метода иногда возникают некоторые трудности: сглаживая тот или иной ряд данных, скользящее среднее приводит к смещению значений, соответствующих достаточно гладким кривым. Для выявления этого эффекта рассмотрим, на-

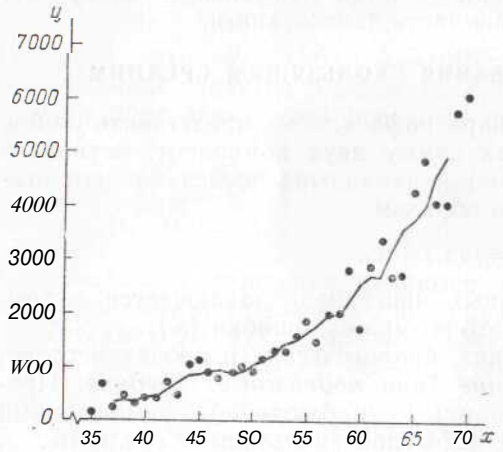


Рис. 4.2.1. Сглаживающий эффект скользящего среднего

пример, кривую квадратичной зависимости $y = 1100 + 2x - 5x^2$. Численные значения этой функции приведены в табл. 4.2.2. Сглаженные значения были получены применением скользящего среднего из пяти последовательных значений, смещение очевидно: каждое численное значение уменьшилось на 10.

Таблица 4.2.2. **Смещение, вызванное сглаживанием простым скользящим средним**

x	Наблюдаемое значение y	Сглаженное значение Y	x	Наблюдаемое значение y	Сглаженное значение Y
0	1100		5	985	975
1	1097		6	932	922
2	1084	1074	7	869	859
3	1061	1051	8	796	
4	1028	1018	9	713	

Итак, отметим три свойства скользящих средних:

1. Они уменьшают нерегулярность колебания в ряде.
2. Они смещают сглаженные значения [1].
3. Они не дают начальные и конечные значения ряда (концы таблицы).

Сглаженные данные, приведенные на рис. 4.2.1, могут быть сделаны еще более гладкими при повторном применении метода скользящего среднего. При этом произойдет дальнейшая потеря данных в начале и конце таблицы и возрастет опасность смещения.

Нет необходимости ограничиваться формулами простых средних, можно использовать *взвешенные средние*. Например, вместо приведенной пятичленной формулы простого среднего можно применить пятичленную формулу взвешенного среднего:

$$Y_i = \frac{1}{9}(y_{i-2} + 2y_{i-1} + 3y_i + 2y_{i+1} + y_{i+2}).$$

Не обязательно настаивать на том, чтобы все веса были положительны. Можно разработать формулы такого сглаживания, при котором не происходит искажений в значениях многочленов третьей (или меньшей) степени, подобные формулы содержат отрицательные члены. Большинство гладких функций может быть успешно представлено наборами отрезков графиков многочленов невысокой степени, при использовании подобных «неискажающих» формул опасность смещений мала. Для заполнения концов таблицы можно применять несимметричные формулы, обладающие теми же свойствами.

Запишем y_0 вместо y_i, y_1 вместо y_{i+1}, 02 вместо y_{i+2} и т. д. Гильдеоранд [43] приводит следующие неискажающие пятичленные формулы:

$$Y_0 = \frac{1}{35}(-3y_{-2} + 12y_{-1} + 17y_0 + 12y_1 - 3y_2), \qquad (4.2.1)$$

$$Y_1 = \frac{1}{35}(2y_{-2} - 8y_{-1} + 12y_0 + 27y_1 + 2y_2), \qquad (4.2.2)$$

$$Y_2 = \frac{1}{70}(-y_{-2} + 4y_{-1} - 6y_0 + 4y_1 + 69y_2). \qquad (4.2.3)$$

[1] Метод уменьшения смещений описан в примере 7.3.1.

Формула (4.2.1) предназначена для основной части таблицы, а (4.2.2) и (4.2.3) — для подсчета двух последних значений таблицы. Формулу для Y_{-1} выводят из (4.2.2) при записи y_2 вместо y_{-2}, y_1 вместо y_{-1}, y_{-1} вместо y_1 и y_{-2} вместо y_2. Формулу для Y_{-2} получают заменой индексов в (4.2.3).

С помощью более длинных формул, вообще говоря, можно добиться большего сглаживающего эффекта. Гильдебранд предлагает следующие семичленные формулы:

$$\left.\begin{aligned} Y_0 &= \frac{1}{21}(-2y_{-3}+3y_{-2}+6y_{-1}+7y_0+6y_1+3y_2-2y_3)\\ Y_1 &= \frac{1}{42}(y_{-3}-4y_{-2}+2y_{-1}+12y_0+19y_1+16y_2-4y_3)\\ Y_2 &= \frac{1}{42}(4y_{-3}-7y_{-2}-4y_{-1}+6y_0+16y_1+19y_2+8y_3)\\ Y_3 &= \frac{1}{42}(-2y_{-3}+4y_{-2}+y_{-1}-4y_0-4y_1+8y_2+39y_3) \end{aligned}\right\}. \qquad (4.2.4)$$

Были разработаны также неискажающие формулы с оптимальными сглаживающими свойствами, и читателю можно рекомендовать их вместо формул, приведенных выше, при сглаживании значений, которые расположены в основной части таблицы (см. [67, с. 68—72]). Веса для оптимальных сглаживающих формул с числом членов 5, 7, 9, 11, 13, 15, 17, 19, 21 и 23 приведены в табл. 4.2.3. Формулы симметричны, коэффициент K_0 соответствует центральному «весу», а K_1 — каждому из весов, примыкающих к центральному, и т. д.

Оптимальные формулы сглаживания для заполнения концов таблицы были предложены Гревиллом (см. [35], [36]). Однако для этой цели часто вполне пригодны приведенные выше несимметричные пятичленные и семичленные формулы; дополнительное сглаживание может быть получено повторным применением процедуры сглаживания.

Пример 4.2.1. Применим оптимальную девятичленную сглаживающую формулу из табл. 4.2.3 к исходным данным табл. 4.2.1.

Эта задача легко может быть решена на среднем по размеру программируемом калькуляторе, вручную же решать ее утомительно. Например, сглаженное значение в точке $x = 39$ вычисляется следующим образом:

$$\begin{aligned} Y_{39} = {} & -0{,}040\,724 \times 95 - 0{,}009\,873 \times 638 + 0{,}118\,470 \times 191 + \\ & + 0{,}266\,557 \times 419 + 0{,}331\,140 \times 278 + 0{,}266\,557 \times 381 + \\ & + 0{,}118\,470 \times 384 - 0{,}009\,873 \times 665 - 0{,}040\,724 \times 477 = 337. \end{aligned}$$

Та же процедура применяется для сглаживания значений при $x = 40, 41, \ldots, 66$. Оптимальная семичленная формула может быть использована для получения сглаженных значений при $x=38$ и $x = 67$. Сглаженные значения в остальных точках можно получить с помощью формулы (4.2.4). Результаты этих вычислений приве-

Таблица 4.2.3. **Коэффициенты оптимально сглаживающих формул скользящего среднего**

	Число членов в формуле				
	5	7	9	11	13
K_0	0,559 441	0,412587	0,331 140	0,277 945	0,240 057
K_1	0,293 706	0,293 706	0,266 557	0,238 693	0,214 337
K_2	−0,073 427	0,058741	0,118 470	0,141 267	0,147 356
K_3		−0,058 741	-0,009873	0,035 723	0.065 492
K_4			-0,040 724	−0,026 792	0,000 000
K_5				−0,027 864	-0,027 864
K_6					−0,019 350
K_7					
K_8					
K_9					
K_{10}					
K_{11}					

	Число членов в формуле				
	15	17	19	21	23
K_0	0,211 541	0,189 231	0,171 266	0,156 469	0,144 060
Ал	0,193 742	0,176 390	0,161 691	0,149 136	0,138318
K_2	0,145 904	0,141 112	0,134 965	0,128423	0,121 949
K_3	0,082918	0,092 293	0,096 658	0,097 956	0,097 395
K_4	0,024 027	0,042 093	0,054 685	0,063 038	0,068 303
K_5	−0,014 134	0,002 467	0,017 475	0,029 628	0,038 933
K_6	−0,024 499	-0,018640	-0,008 155	0,003 119	0,013 430
K_7	-0,013730	−0,020 370	−0,018 972	-0,012896	-0,004948
K_8		−0,009 960	−0,016 601	−0,017 614	-0,014527
K_9			-0.007 378	−0,013 455	−0,015 687
K_{10}				-0,005570	-0,010918
K_{11}					-0,004 278

Замечание. Формула для значения K_r в $(2m-3)$-членном скользящем среднем имеет вид

$$K_r = \frac{315\,\{(m-1)2 - r^2\}\,\{m^2 - r^2\}\,\{(m+1)^2 - r^2\}\{(3m^2 - 16) - 1\,1r^2\}}{8m\,(m^2 - \backslash)\,(4m^2-1)\,(4m^2-9)\,(4m^2-25)}$$

Таблица 4.2.4. **Сглаживающий эффект девятичленной формулы**

x	Наблюдаемое значение y	Сглаженное значение Y	x	Наблюдаемое значение y	Сглаженное значение Y
35	95	176	53	1274	1397
36	638	386	54	1579	1542
37	191	426	55	1873	1631
38	419	370	55	1482	1786
39	278	337	57	2002	1947
40	381	357	58	2011	2083
41	384	429	59	2827	2281
42	665	549	60	1722	2528
43	477	725	61	2942	2687
44	1015	973	62	3374	2805
45	1093	941	63	2735	3045
45	860	930	(A	2731	3452
47	779	876	65	4323	3853
48	862	835	66	4926	4219
49	951	882	67	4114	4426
50	866	971	68	4099	4559
51	1109	1074	69	5835	5022
52	1291	1229	70	6103	6395

дены в табл. 4.2.4. На рис. 4.2.2 отчетливо видны результаты сглаживания. Процедуру можно повторить для достижения еще большей сглаженности.

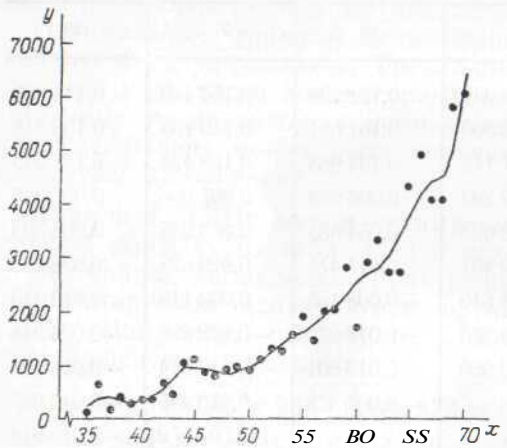


Рис. 4.2.2. Сглаживающий эффект **девятичленной** формулы

Литература: [35], [36], [41, с. 247—251], [43, с. 295—302], [67, с. 25—33, 68—72], [73, с. 287—293], [96, с. 176—180].

4.3. СПЛАЙНЫ (СПЛАЙН-ФУНКЦИИ)

Всегда можно подобрать такой многочлен, кривая которого проходит через n заданных точек (разумеется, при условии, что никакие две из них не лежат на одной и той же вертикали). Однако в общем случае порядок такого многочлена равен $n-1$, и если n велико, то при росте аргумента характер изменения значений задаваемой им функции будет отчетливо волнообразным. Но виду подобную кривую трудно признать «сглаженной».

Часто[2] эту задачу удается решить с помощью *сплайна* (слово «сплайн» обозначает буквально чертежное устройство, представлтощее собой полосу из некоторого гибкого материала, к которой в определенных точках могут прикрепляться грузики; они должны обеспечить прохождение искомой кривой через некоторые заданные точки или вблизи этих точек).

Сплайном (или кусочно-сопряженной функцией) называют и вычислительной математике такую функцию, кривая которой состоит из отрезков полиномиальных кривых; эти отрезки состыкованы так, что производные полученной функции (до порядка на единицу меньшей степени используемых полиномов) непрерывны на всем рассматриваемом промежутке. Подобные функции \ тобны для интерполяции. При многочленах относительно низкой степени часто можно избежать явно волнообразного поведения функции, которое характерно для случаев сглаживания большого числа эмпирических наблюдений единственным многочленом. С другой стороны, данная процедура обеспечивает гораздо большую "гладкость, чем традиционная кусочно-линейная интерполяция, при которой интерполяционная функция имеет разрывы даже в первой производной. Сплайн очевидным образом обеспечивает непрерывность производных интерполяционной функции до максимально высокого возможного порядка при выполнении условия, что степень многочленов, используемых для сглаживания исходных данных, ниже степени того единственного многочлена, кри-пая которого проходит через все заданные точки.

Рассмотрим множество точек на плоскости, координаты которых соответствуют рассматриваемым данным (a, $f(a)$), $(b, f(b))$, (г, $f(c)$), $(d, f(d))$, и предположим, что уже найден многочлен степени три или ниже, кривая которого проходит через точки $(a, f(a))$ и $(b, f(b))$. Теперь нам нужно подобрать многочлен третьей степени

$$y = A + B(x-b) + C(x-b)^2 + D(x-b)^3, \qquad (4.3.1)$$

кривая которого проходила бы через точки $(b, f(b))$ и $(c, f(c))$. Необходимо определить значения четырех констант A, B, C и D, и значения $f(b)$ и $f(c)$ рассматриваются как два ограничения. Можно также обусловить, чтобы кривая в точке $x = b$ имела такой же наклон и такое же значение второй производной, что и у ранее найденного многочлена. Теперь мы имеем четыре урав-

---

* Речь идет о минимальном порядке многочлена, соответствующего заданным n точкам.— *Примеч. пер.*

'•' По не всегда, см. пример 4.3.2.

нения, которым должны удовлетворять неизвестные A, B, C и D. Легко обнаружить, что

$$f(b) = A,$$
$$f(c) = A + B(c - b) + C(c - b)^2 + D(c - b)^3,$$
$$\text{га} = B,$$
$$\text{f''}(b) = 1C.$$

Таким образом,

$$\left.\begin{aligned} A &= f(b), \\ B &= f'(b), \\ C &= \tfrac{1}{2} f''(b), \\ D &= \{f(c) - A - B(c - b) - C(c - b)^2\}/(c - b)^3 \end{aligned}\right\}. \qquad (4.3.2)$$

Теперь можно осуществить интерполяцию между значениями b и c. Когда этот этап выполнен, мы переходим к построению следующего отрезка кривой, проходящего через значения $f(c)$ и $f(d)$. Для этого необходимо вычислить тангенс угла наклона и значение второй производной нашего многочлена в точке $x = c$, а именно

$$\left.\begin{aligned} f'(c) &= B + 2C(c - b) + 3D(c - b)^2 \\ f''(c) &= 2C + 6D(c - b) \end{aligned}\right\}'. \qquad (4.3.3)$$

Пример 4.3.1. Воспользуемся данными табл. 4.2.1 для демонстрации вычислительной процедуры. Сначала необходимо выбрать опорные точки, или «узлы». На основе графика, приведенного на рис. 4.2.1, предположим, что можно получить гладкую кривую, если взять значения скользящих средних в точках $x = 37$, $x = 42$ и $x = 55$ и исходное значение в точке $x = 69$. Итак, проведем сплайн-кривую через эти точки, а именно

x	y
37	324
42	584
55	1642
69	5835

Представляется удобным отрезок прямой в промежутке между $x = 37$ и $x = 42$. В этом случае тангенс угла наклона в точке $x = 42$ равен $(584 - 324)/5 = 52$, а вторая производная в этой точке равна нулю. Теперь подберем многочлен третьей степени для отрезка между $x = 42$ и $x = 55$. Из формул (4.3.2) получаем:

$$A = 584; \quad B = 52; \quad C = 0;$$
$$D = (1642 - 584 - 52 \times 13 - 0 \times 13^2)/13^3 = 0{,}173\,873\,4.$$

Теперь с помощью формулы (4.3.1) можно выполнить интерполяцию для промежутка между $x = 42$ и $x = 55$.

Для того чтобы определить следующий отрезок кривой, необходимо вычислить тангенс угла наклона и значение второй производной в точке $x = 55$. В соответствии с формулой (4.3.3) получаем:

$$f'(55) = 52 + 2 \times 0 \times 13 + 3 \times 0{,}173\,8734 \times 13^2 = 140{,}153\,81;$$
$$f''(55) = 2 \times 0 + 6 \times 0{,}173\,873\,4 \times 13 = 13{,}562\,125.$$

Итак, для следующего отрезка кривой получены значения:

$$A = 1642; \quad B = 140{,}15381; \quad C = \tfrac{1}{2}(13{,}562\,125) = 6{,}781\,063;$$
$$D = (5835 - 1642 - 140{,}15381 \times 14 - 6{,}781\,063 \times 14^2)/14^3 = 0{,}328\,629\,1.$$

Теперь можно перейти к интерполяции значений в промежутке между 55 и 69. Полностью полученная кривая показана на

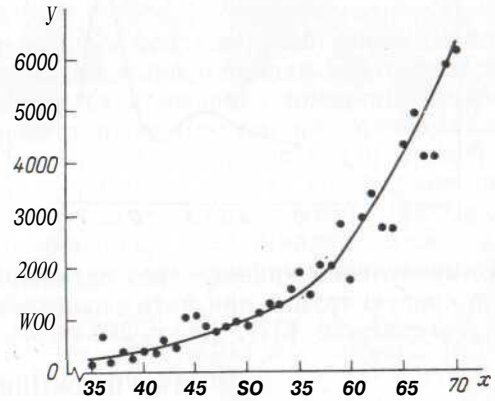


Рис. 1.3.1. Кривая сплайна с узлами в точках $x = 37$, $x = 42$, $x = 55$ и $x = 69$

рис. 4.3.1. Значение в точке $x = 70$ было получено экстраполяцией последнего отрезка кривой, а значения в точках $x = 35$ и $x = 36$ — экстраполяцией начального прямолинейного отрезка. Обратите внимание на гладкость полученной кривой. Округленные значения данного сплайна при целых значениях аргумента в рассматриваемом диапазоне приведены в табл. 4.3.1.

Читателю следует обратить внимание на весьма произвольный выбор опорных точек в данном примере. Некоторые исследователи меняют абсциссы и ординаты рассматриваемых узлов для того, чтобы получить оптимальную (относительно некоторого критерия) кривую. Для реализации этой задачи необходима соответствующая программа.

Пример 4.3.2. Этот пример приведен как предостережение. Значения функции $\ln x$ при целых положительных значениях x представлены графически на рис. 4.3.2. Через первые три точки была проведена парабола и далее был применен метод сплайна

Таблица 4.3.1. **Кривая, полученная с помощью сплайна**

x	$f(x)$	x	$f(x)$	x	$f(x)$
35	220	47	866	59	2332
36	272	48	934	60	2553
37	324	49	1008	61	2798
38	376	50	1089	62	3068
39	428	51	1179	63	3365
40	480	52	1278	64	3692
41	532	53	1387	60	4050
42	584	54	1508	66	4442
43	636	55	1642	67	4868
44	689	56	1789	68	5332
45	745	57	1952	69	5835
46	803	58	2132	70	6379

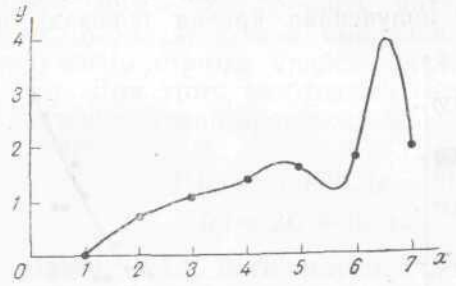


Рис. 4.3.2. Кривая сплайна, проходящая через значения в целочисленных точках $x = 1, 2, \ldots, 7$

для проведения кривой через остальные точки графика. Полученную кривую трудно признать «сглаженной».

Литература: [37], [96, с. 265—267].

4.4. УПРАЖНЕНИЯ

1. Примените формулу (4.2.4) к какому-либо многочлену третьей степени и проверьте утверждение о том, что эти формулы не приводят к смещению значений такого многочлена.
2. Проверьте утверждение о том, что оптимальная пятичленная формула сглаживания из табл 4.2.3 не смещает значения многочлена третьей (степени.
3. Примените метод из примера 4.2.1 для Дальнейшего сглаживания значений, приведенных в табл. 4.2.4.
4. В примере 4 3 1 взяты четыре опорные точки. Через первые три точки проведите параболу методом, описанным в Параграфе 6.10, а далее с помощью сплайн-метода получите всю кривую.

0. ПЛОЩАДЬ ПОД КРИВОЙ

Для получения интеграла какой-нибудь функции приходится прибегать к" численным методам, когда функция задана рядом эмпирических значений, а также когда функция задана аналитически, но ее интеграл не может быть представлен аналитическим выражением. В данной главе приведены формулы численного интегрирования.

5.1. ВВЕДЕНИЕ

Интеграл той или иной функции определяет значение площади фигуры, заключенной между кривой, являющейся графиком рассматриваемой функции, и горизонтальной осью системы координат. Если функция положительна на всем рассматриваемом промежутке, то указанная площадь будет положительна; если функция на всем промежутке отрицательна, то площадь полученной фигуры считают отрицательной. Если функция задана явным аналитическим выражением, то соответствующая площадь может иногда выражаться аналитической зависимостью. Однако в огромном большинстве случаев подобного аналитического выражения для площади подобрать нельзя (например, для площади, ограниченной сверху нормальной кривой ошибок). В таких ситуациях следует применять какой-либо из методов численною интегрирования; подобная необходимость возникает и тогда, когда функция задана таблицей эмпирических значений, а не математической формулой. В данной главе приведены формулы численного интегрирования.

Отметим, что при использовании той или иной формулы численного интегрирования для нахождения площади под некоторой кривой на самом деле оценивается суммарная площадь, ограниченная сверху рядом отрезков аппроксимирующих полиномиальных дуг. Эти отрезки частично лежат выше истинной кривой и частично расположены ниже ее. Тем не менее получаемая при этом оценка площади является достаточно точной, поскольку интегрирование (как и суммирование) представляет собой сглаживающую процедуру, нейтрализующую подобные отклонения на отдельных участках кривой.

5.2. ФОРМУЛА ТРАПЕЦИЙ

Если график функции $f(x)$ на интервале $(0,1)$ — отрезок прямой линии, то

$$\int_0^1 f(x)\,dx = \frac{1}{2}\{f(0) + f(1)\}.$$

Доказать это совсем просто. При применении этой формулы для нахождения значения интеграла произвольной функции, заданной на указанном интервале, полученная оценка будет весьма далека от истинного значения, если" график рассматриваемой функции существенно отклоняется от соответствующего прямолинейного отрезка. Для начального интервала длины h формула имеет вид

$$\int_0^h f(x)\,dx = \frac{1}{2} h\{f(0) + f(h)\}. \qquad (5.2.1)$$

Если рассматриваемый интервал имеет длину nh, то в формулу трапеций входят соответственно $n + 1$ ординат:

$$\int_0^{nh} f(x)\,dx = \frac{1}{2} A\{f(0) + 2f(A) + 2f(2A) + 2f((n-1)h) + f(nh)\}. \quad (5.2.2)$$

В случае когда кривая, являющаяся графиком положительной функции, выпукла вниз, формула трапеций дает завышенную оценку площади; если она выпукла вверх, то оценка занижена. Для получения точных результатов значение h должно быть достаточно малым.

Пример 5.2.1. С помощью формулы трапеций вычислим площадь под кривой $y = \exp(1/x)$ между вертикалями $x = 1$ и $x = 2$ с точностью до четырех значащих цифр.

Начнем с выбора значения $h = 1$ и возьмем две соответствующие ординаты. Интеграл приближенно равен:

$$\frac{1}{2}(e^{1/1} + e^{1/2}) = 2,183\,50.$$

В процессе повторного вычисления при $h = 1/2$ и трех ординатах получаем уточненное приближение:

$$\frac{0,5}{2}(e^{1/1} + 2e^{1/1,5} + e^{1/2}) = 2,065\,62.$$

Повторяя последовательно вычисления для уменьшенного каждый раз вдвое значения h, имеем следующие результаты:

$h = \frac{1}{4}$ (5 ординат); площадь $\doteq$ 2,031 89;

$h = \frac{1}{8}$ (9 ординат); площадь = 2,023 05;

$h = \frac{1}{16}$ (17 ординат); площадь = 2,020 81;

$h = \frac{1}{32}$ (33 ординаты); площадь = 2,020 25;

$h = \frac{1}{64}$ (65 ординат); площадь = 2,020 11.

Два последних приближения равны друг другу с точностью до четырех значащих цифр; отсюда заключаем, что площадь под рассматриваемой кривой равна 2,020 с точностью до четырех значащих цифр. Заметьте, что кривая выпукла вверх и значения полученных последовательных оценок монотонно убывают. Для

вычисления площади с заданной точностью нам потребовалось большое число ординат.

Литература: [17, с. 119—126], [43, с. 73], [100, с. 156—158, русский перевод с. 150—151].

5.3. ФОРМУЛА СИМПСОНА

Формула трапеций может обеспечить приемлемую точность вычисления интеграла лишь в том случае, когда ординаты взяты в очень близких точках. Это требует выполнения большого числа арифметических операций. Возникает проблема поиска альтернативных формул интегрирования.

Предположим, что $f(x)$ — многочлен третьей степени. Нетрудно доказать, что в этом случае

$$\int_{-h}^{h} f(x)\,dx = \frac{1}{3} A\{f(-A) + 4f(0) + f(h)\}. \quad (5.3.1)$$

Это равенство называется формулой Симпсона (для трех точек), ее применяют гораздо чаще, чем формулу трапеций. Когда рассматриваемая функция в пределах интегрирования может быть точно представлена многочленом третьей степени, формула Симпсона всегда дает точное значение интеграла. В общем случае (для $2n + 1$ точек) она имеет вид

$$\int_0^{2nh} f(x)\,dx = \frac{1}{3} h\{f(0) + 4f(h) + 2f(2h) + 4f(3h) + \ldots$$
$$\ldots + 4f((2n-1)A) + f(2nh)\}. \quad (5.3.2)$$

Пример 5.3.1. Воспользуемся формулой Симпсона для вычисления площади под кривой $y = \exp(1/x)$ между вертикалями $x = 1$ и $x = 2$ с точностью до четырех значащих цифр. Начнем со значения $h = 0,5$ и трех ординат.

$h = 0,5$

t	$\exp\{1/(1+t)\}$	Множитель Симпсона	Произведение
0,0	2,718 281 8	1	2,718 281 8
0,5	1,947 734 0	4	7,790 936 0
1,0	1,648 721 3	1	1,648 721 3
			Всего 12,157 939 1

Искомая площадь приближенно равна $\frac{1}{3}(0,5)(12,157\,939\,1) = 2,026\,32$. Теперь повторим вычисления с $h = 0,25$ и пятью ординатами.

$h = 0{,}25$

t	$\exp\{1/(1+t)\}$	Множитель Симпсона	Произведение
0,00	2,7182818	1	2,718 281 8
0,25	2,225 540 9	4	8,902 163 6
0,50	1,947 734 0	2	3,895 468 0
0,75	1,770 795 0	4	7,083 180 0
1,00	1,648 721 3	1	1,648 721 3
			Всего 24,247 814 7

Площадь приближенно равна $\frac{1}{3}(0{,}25)(24{,}247\,814\,7) = 2{,}020\,65$.

Продолжая последовательно уменьшать вдвое длину интервала, получаем следующие результаты:

$h = 0{,}125$ (9 ординат); площадь $= 2{,}020\,10$;

$h = 0{,}0625$ (17 ординат); площадь $= 2{,}020\,06$.

Последние два приближения равны между собой с точностью до четырех значащих цифр, отсюда заключаем, что искомая площадь под кривой равна 2,020 с точностью до четырех значащих цифр. Обратите внимание, что этот результат достигнут при относительно небольшом объеме вычислений (по сравнению с вычислениями по формуле трапеций).

Литература: [17, с. 134—136], [27, с. 174—180], [43, с. 73—75, 141, 145—146], [73, с. 118—120], [100, с. 156—158; русский перевод с. 150—151].

5.4. ФОРМУЛА ТРЕХ ВОСЬМЫХ

Применение для численного интегрирования формулы Симпсона возможно во всех случаях, когда заданный промежуток разбит на четное число равных отрезков и рассматривается нечетное число ординат, соответствующих концам этих отрезков. Однако во многих экспериментальных ситуациях число оцененных ординат четно, и тогда нельзя непосредственно применять формулу Симпсона. Формула трех восьмых использует четыре ординаты. Сочетая ее с формулой Симпсона, можно производить численное интегрирование данного типа для любого числа равных отрезков.

Предположим вновь, что $f(x)$ — многочлен третьей степени. Нетрудно доказать, что

$$\int_0^{3h} f(x)\,dx = \frac{4}{3} * \{(f(0) + 3/(h) + 3/(Щ + f(3h)\}. \qquad (5.4.1)$$

Обобщение этой формулы на случай произвольного нечетного числа отрезков очевидно.

Пример 5.4.1. Некоторая гладкая кривая $y = f(x)$ проходит через следующие точки: (1; 3,724), (2; 4,492), (3; 5,359), (4; 6,332), (5; 7,415) и (6; 8,615). Найдите площадь под этой кривой между вертикалями $x = 1$ и Л" = 6.

К шести ординатам нельзя применить формулу Симпсона, а также нельзя непосредственно использовать и формулу трех восьмых. Можно было бы воспользоваться формулой трапеций, но она не дает достаточно точного результата. Однако не все потеряно, поскольку можно интегрировать в промежутке от $x = 1$ до $x = 3$ по формуле Симпсона, а в промежутке от $x = 3$ до $x = 6$ — по формуле трех восьмых. Эти вычисления можно произвести так:

$h = 1$

x	$f(x)$	Множитель Симпсона	Произведение
1	3,724	1	3,724
2	4,492	4	17,968
3	5,359	1	5,359
			Всего 27,051

Площадь под кривой между $x = 1$ и $x = 3$ равна приблизительно $\frac{1}{3}(27{,}051) = 9{,}017$.

$h = 1$

x	$f(x)$	Множитель трех восьмых	Произведение
3	5,359	1	5,359
4	6,332	3	18,996
5	7,415	3	22,245
6	8,615	1	8,615
			Всего 55,215

Площадь под кривой между $x = 3$ и $x = 6$ равна приблизительно $\frac{3}{8}(55{,}215) = 20{,}706$. Общая площадь между $x = 1$ и $x = 6$ равна, следовательно, 29,723.

Литература: [27, с. 181], [43, с. 73], [100, с. 156—158; русский перевод с. 150—151].

5.5. ДРУГИЕ МЕТОДЫ ЧИСЛЕННОГО ИНТЕГРИРОВАНИЯ, ФОРМУЛА УЭДДЛА

Формула Симпсона и формула трех восьмых пригодны для многих целей. Однако существует и много других формул интегрирования. Следует упомянуть, в частности, формулу Уэддла:

$$\int_0^{6h} f(x)\,dx = \frac{3}{10}A\{f(0) + 5f(A) + f(2A) + 6f(3A) + f(4h) + 5f(5h) + f(6h)\}. \qquad (5.5.1)$$

Эта формула точна, если на всем промежутке интегрирования $f(x)$ является многочленом пятой (или меньшей) степени. Легко получить общий вид этой формулы.

Пример 5.5.1. Гладкая кривая $y = f(x)$ проходит через точки: (0; 3,049), (1; 3,724), (2; 4,492), (3; 5,359), (4; 6,332), (5; 7,415) и (6; 8,615). Применим формулу Уэддла, чтобы найти площадь под кривой между вертикалями $x = 0$ и $x = 6$.

Расчеты можно описать следующим образом:

$h = 1$

x	$f(x)$	Множитель Уэддла	Произведение
0	3,049	1	3,049
1	3,724	5	18,620
2	4,492	1	4,492
3	5,359	6	32,154
4	6,332	1	6,332
5	7,415	5	37,075
6	8,615	1	8,615
			Всего 110,337

Искомая площадь приблизительно равна $\frac{3}{10}$ (1) (110,337) = = 33.101.

Литература: [17, с. 137], [27, с. 182—185], [43, с. 73, 160—161], [73, с. 118—120], [92, с. 554—558].

5.6. ОРДИНАТЫ В НЕРАВНООТСТОЯЩИХ ТОЧКАХ

Часто в экспериментах приходится сталкиваться со случаями, когда ординаты заданы в точках, расположенных не через равные промежутки. В подобных ситуациях можно применять формулу трапеций, однако при этом полученный результат вряд ли будет очень точен. Иногда можно применять и формулу Симпсона, и формулу трех восьмых. Например, если ординаты соответствуют точкам $x = 1$; 1,5; 2; 2,5; 3; 4; 5 и 6, то могут оказаться

полезными обобщенная формула Симпсона в промежутке от $x = 1$ до $x = 3$ и формула трех восьмых в промежутке от $x = 3$ до $x = 6$.

Однако гораздо чаще необходим другой подход. Можно оценить значения ординат в равноотстоящих точках с помощью методов интерполяции (см. гл. 6) или сплайн-метода (см. параграф 4.3). После этого можно воспользоваться формулой Симпсона или формулой трех восьмых. Если для интерполяции применяется сплайн-метод, то тем самым задается некоторый набор отрезков полиномиальных дуг. В этом случае искомую площадь, вероятно, лучше вычислить, непосредственно интегрируя (на соответствующем промежутке) каждый из этих полиномов. Все это следует выполнять с большой осторожностью, так как полученная кривая может обладать нежелательными излишними колебаниями. Поэтому рекомендуется сделать набросок графика интерполирующей функции, рассмотрев ряд ее промежуточных значений. Пример 4.3.2 является предостережением.

Пример 5.6.1. Гладкая кривая проходит через точки (0; 1,0000), (0,5; 0,7788), (2; 0,3679), (3; 0,2231) и (5; 0,0821). Найдем площадь под этой кривой между вертикалями $x = 0$ и $x = 5$.

С помощью метода интерполяции (см. параграф 6.10) находим, что через первые три из указанных точек проходит следующая парабола:

$$y = 1{,}0000 - 0{,}484\,516x + 0{,}08423x^2.$$

Площадь под этой параболой между значениями $x = 0$ и $x = 2$ **равна:**

$$[x - 0{,}242\,258x^2 + 0{,}02807x^3]_0^2 = 1{,}2556.$$

Можно воспользоваться сплайн-методом, описанным в параграфе 4.3, для того, чтобы провести кубическую кривую через точки (2; 0,3679) и (3; 0,2231):

$$y = 0{,}3679 - 0{,}147\,583(x - 2) + 0{,}08423(x - 2)^2 - 0{,}08145(x - 2)^3.$$

Сделав замену $X = x - 2$, определим площадь под этой кубической кривой между $x = 2$ и $x = 3$:

$$\left[0{,}3679X - 0{,}147583\left(\frac{1}{2}X^2\right) + 0{,}08423\left(\frac{1}{3}X^3\right) - 0{,}08145\left(\frac{1}{4}X^4\right)\right]_0^1 = 0{,}3018.$$

Отрезок сплайна, проходящий через две последние точки, задается уравнением

$$y = 0{,}2231 - 0{,}223\,46(x - 3) + 0{,}160\,116(x - 3)^2 - 0{,}041\,816(x - 3)^3.$$

Соответствующая площадь под этой кривой равна 0,2590.

Отсюда заключаем, что общая площадь под рассматриваемой кривой (уравнение которой нам не известно) между вертикалями

$x = 0$ и $x = 5$ приблизительно равна $1{,}2556 + 0{,}3018 + 0{,}2590$, или 1,82 с точностью до трех значащих цифр. На самом деле в нашем примере заданные точки лежат на кривой $y = \exp\left(-\frac{1}{2}x\right)$, и площадь под этой кривой, как известно, равна 1,84 с точностью до трех значащих цифр. Таким образом, полученному ответу соответствует ошибка в 1 %.

Следует обратить внимание на резко возросший объем вычислений при определении площади под кривой в тех случаях, когда невозможно применять обычные формулы.

5.7. УПРАЖНЕНИЯ

1. С помощью формулы трапеций найдите площадь под кривой, описанной в примере 5.6.1.
2. С помощью формулы Симпсона найдите площадь под кривой, описанной в примере 5.5.1.
3. С помощью формулы трех восьмых найдите площадь под кривой, описанной в примере 5.5.1.
4. Воспользуйтесь двумя параболами, одна из которых проходит через первые три точки, а вторая — через последние три точки, для оценки площади под кривой, описанной в примере 5.6.1.
5. Необходимо сравнять некоторый холм, чтобы подготовить спортивную площадку. Высота поверхности холма над окончательным уровнем площадки в точке (x, y) задается значением $f(x, y)$. Зная, что площадка должна иметь размеры $2h \times 2k$, докажите, что общий объем перемещенного со склонов грунта будет приблизительно равен:

$$\frac{1}{9}hk\{(f_{0,0} + f_{0,2} + f_{2,0} + f_{2,2}) + 4(f_{0,1} + f_{1,0} + f_{1,2} + f_{2,1}) + 16f_{1,1}\},$$

где $f_{r,s} = f(rh, sk)$.

Указание. Воспользуйтесь формулой Симпсона.

6. КОНЕЧНЫЕ РАЗНОСТИ, ИНТЕРПОЛЯЦИИ И ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

В данной главе изучаются свойства конечных разностей и демонстрируется исчисление конечных разностей для численного дифференцирования и интерполяции. Обычные методы конечных разностей непригодны в тех ситуациях, когда ординаты соответствуют неравноотстоящим точкам. Для решения этой проблемы в параграфе 6.7 вводятся разделенные разности. Простые методы нахождения прямой, проходящей через две точки, и параболы, проходящей через три точки, обсуждаются в параграфах 6.9 и 6.10.

6.1. СПОСОБ ПОСТРОЕНИЯ КОНЕЧНЫХ РАЗНОСТЕЙ

Тангенс угла наклона кривой $y = f(x)$ в заданной точке x задается выражением

$$\frac{\Delta y}{dx} = f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

Этот тангенс угла наклона (или *производная y* по *x)* обычно легко может быть вычислен в тех случаях, когда аналитическое выражение функции *f(x)* известно. Однако во многих экспериментах удается получить некоторую гладкую кривую, которая, вообще говоря, не может быть задана каким-либо известным аналитическим выражением. В подобных случаях скорее можно найти некоторое приближение тангенса угла наклона в точке *x*, используя выражение $\{f(x+h) - f(x)\}$, чем предел этого выражения при $h \to 0$ (рис. 6.1.1). Разность $f(x + h) - f(x)$ называется *конечной разностью.*

Даже при известном аналитическом выражении, задающем функцию $f(x)$, метод дифференцирования с помощью конечных

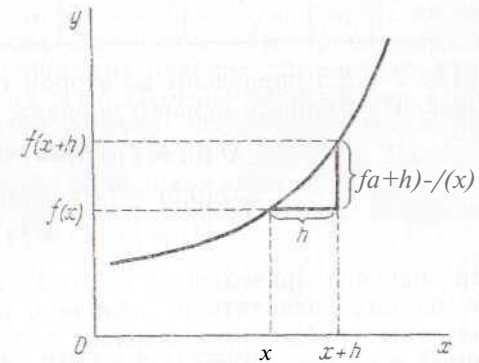


Рис. 6.1.1. Конечно-разностное приближение тангенса угла наклона кривой

разностей может оказаться предпочтительнее, чем аналитические формулы дифференцирования (вследствие громоздкого выражения для производной; см. упражнение 2 в конце главы).

Методы конечных разностей применимы также для получения численных решений дифференциальных уравнений, при этом производные в этих уравнениях замещаются приближенно равными им конечными разностями.

Другой областью применения этих методов является *интерполяция.* При составлении таблицы значений некоторой непрерывной функции необходимо стремиться к максимальной компактности, что обеспечит получение значений рассматриваемой функции с заданной степенью точности. Методы конечных разностей позволяют «читать между строк таблицы» и интерполировать.

6.2. ТАБЛИЦА РАЗНОСТЕЙ

Процедуры расчета конечных разностей при определении значений многочленов всегда приводят к точным результатам. Рассмотрим таблицу разностей для многочлена третьей степени $f(x) = 8x^3 - 2x + 1$. Значения этой функции в точках $x = 0$; 0,5;

Таблица 6.2.1. **Таблица разностей**

x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$
0,0	1				
		0			
0,5	1		6		
		6		6	
1,0	7		12		0
		18		6	
1,5	25		18		0
		36		6	
2,0	61		24		
		60			
2,5	121				

1; 1,5; 2 и 2,5 приведены во втором столбце табл. 6.2.1. Нетрудно вычислить разности первого порядка:

$$\Delta f(0) = f(0{,}5) - f(0),$$
$$\Delta f(0{,}5) = f(1) - f(0{,}5),$$
$$\Delta f(1) = f(1{,}5) - f(1) \text{ и т. д.}$$

Эти значения приведены в третьем столбце табл. 6.2.1. Если каждое из них разделить на *величину шага,* равную 0,5, то будут получены грубые приближения тангенса угла наклона данной кривой в различных точках.

В общем случае, когда шаг таблицы равен Л, разность первого порядка функции $f(x)$ в точке $x = a$ определяется следующим образом:

$$\Delta f(a) = f(a + K) - f(a). \qquad (6.2.1)$$

В дифференциальном исчислении часто используется вторая производная (тангенс угла наклона кривой, соответствующей значениям тангенса угла наклона графика исходной функции). Аналогично в исчислении разностей рассматривают разности второго порядка (разности значений разностей первого порядка). Так, для исследуемого многочлена

$$\Delta^2 f(0) = \Delta(\Delta f(0{,}5) - \Delta f(0)) = (f(1) - f(0{,}5)) - (f(0{,}5) - f(0)),$$
$$\Delta^2 f(0{,}5) = \Delta(f(1) - f(0{,}5)) = (f(1{,}5) - f(1)) - (f(1) - f(0{,}5)).$$

Соответствующие численные значения приведены в четвертом столбце табл. 6.2.1. В общем случае для шага таблицы

$$\Delta^2 f(a) = \Delta(f(a + h) - f(a)) = (f(a + 2h) - f(a + h)) - (f(a + h) - f(a)). \qquad (6.2.2)$$

Обобщение на случай разностей более высокого порядка очевидно.

Заметим, что в таблице значения разностей соответствующего порядка расположены в своем столбце на последовательных уровнях, лежащих посередине между уровнями, на которых в предыдущем столбце располагаются величины, исходные для этих разностей. Значение $\Delta^n f(a)$ попадает в столбец разностей n-го порядка на уровень, лежащий на n полуинтервалов ниже, чем уровень значения $f(a)$*. Так, например, $\Delta^2 f(1{,}5)$ «падает» в столбце разностей второго порядка на один полный интервал относительно уровня $f(1{,}5)$. Табл. 6.2.1 показывает, что на этом месте расположено значение 24.

В табл. 6.2.1 видно, что все разности четвертого и более высоких порядков равны нулю, а разности третьего порядка постоянны. Обобщение этого факта очевидно, а именно для многочлена j-й степени разности j-го порядка равны между собой, а разности более высокого порядка равны нулю. Для функции, не являющейся многочленом, при последовательном нарастании порядка разностей обычно сначала происходит падение абсолютных величин этих разностей, а затем их рост. Переход от убывания к возрастанию абсолютных величин обычно происходит после рассмотрения разностей приблизительно третьего порядка. Разности высокого порядка, как правило, не используются в расчетах, это соответствует приближению функции многочленами невысоких степеней.

Литература: [17, с. 82—87], [27, с. 1—6], [41, с. 36—48], [43, с. 91—94], [73, с. 27-30], [84, с. 46—52], [92, с. 527—528].

6.3. ПРОВЕРКА ЧИСЕЛ В ТАБЛИЦЕ

Таблица разностей может быть полезным инструментом при проверке чисел в некоторой таблице значений той или иной функции. Если известно, что табулируемая функция является многочленом, но неизвестна его степень, то можно подвергнуть анализу разности значений функции. Предположим, что имеется развернутая таблица разностей и что все разности четвертого порядка, кроме пятой, равны нулю. Тогда можно прийти к заключению, что табулированная функция является многочленом третьей степени и что одно из значений функции в данной таблице ошибочно. Вследствие этой ошибки в столбце разностей первого порядка содержатся два искаженных значения, в столбце разностей второго порядка — три, в столбце разностей третьего порядка — четыре и в столбце разностей четвертого порядка — пять искаженных значений. Ошибки в таблице разностей распространяются «веером» от исходной ошибки. Значения этих ошибок пропорциональны коэффициентам разложения по формуле бинома[1] $(1 - x)^n$, где n — порядок рассматриваемой разности (см. табл. 6.3.1). Ошибку теперь нетрудно найти и исправить.

* Интервал равен разности высот двух последовательных уровней в столбце.— *Примеч. пер.*

1 См. табл. 1.5.1.

Таблица 6.3.1. **Распространение «веером» ошибок в таблице разностей (на примере многочлена)**

x	x^3	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$
0	0				
		1			
1	1		6		
		7		6	
2	8		12		0
		19		6	
3	27		18		$0+\epsilon$
		37		$6+\epsilon$	
4	64		$24+\epsilon$		$0-4\epsilon$
		$61+\epsilon$		$6-3\epsilon$	
5	$125+\epsilon$		$30-2\epsilon$		$0+6\epsilon$
		$91-\epsilon$		$6+3\epsilon$	
6	216		$36+\epsilon$		$0-4\epsilon$
		127		$6-\epsilon$	
7	343		42		$0+\epsilon$
		169		6	
8	512		48		0
		217		6	
9	729		54		
		271			
10	1000				

Если рассматриваемая функция не является многочленом, то разности высших порядков не равны нулю, но они, как правило, малы и значения их обычно изменяются закономерно. В этих условиях можно все же довольно легко обнаружить ошибку и исправить ее.

Пример 6.3.1. Некоторые из следующих двадцати последовательных значений функции, соответствующих равноотстоящим значениям аргумента, искажены вследствие обычных ошибок при переписывании (см. параграф 2.4). Выявим ошибки и исправим их.

1,7278	4,8818	7,9779	11,2630
2,3424	5,4440	8,6249	11,9398
2,9585	6,0723	9,2752	12,6246
3,5764	6,7041	9,9318	13,3180
4,1964	7,3398	10,5937	14,0205

Начинаем с формирования таблицы разностей (см. табл. 6.3.2; обратите внимание на то, как была упрощена таблица отбрасыванием запятой при записи разностей). На основании анализа таблицы можно предположить, что были искажены три значения. Если ввести соответствующие поправки в эти три

Таблица **6.3.2. Использование таблицы разностей для выявления и исправления ошибок**

$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$
1,7278			
	6146		
2,3424		15	
	6161		3
2,9585		18	
	6179		3
3,5764		21	
	6200		633
4,1964		654	
	6854		−1886
4,8818		−1232	
	5622		1893
5,4440		661	
	6283		−626
6,0723		35	
	6318		4
6,7041		39	
	6357		−15
7,3398		24	
	6381		65
7,9779		89	
	6470		−56
8,6249		33	
	6503		30
9,2752		63	
	6566		−10
9,9318		53	
	6619		21
10,5937		74	
	6693		1
11,2630		75	
	6768		5
11,9398		80	
	6848		6
12,6246		86	
	6934		6
13,3180		92	
	7026		
14,0206			

клетки, то в столбце разностей третьего порядка будет наблюдаться упорядоченное возрастание значений от 3 до 6.

Первая из разностей третьего порядка внутри «веера», исходящего из значения функции 4,8818, имеет значение 633, оно слишком велико и, по-видимому, превосходит истинное значение на 630. В таком случае приведенное значение 4,8818 превосходит истинное значение функции в этой точке на 630×10^{-4}. Отсюда заключаем, что истинное значение в этой клетке равно 4,8188. Тогда рассмотренные разности третьего порядка равны 3, 4, 3 и 4 соответственно.

Литература: [41, с. 44—47], [43, с. 110—112], [73, с 38—40], [84, с. 50—52].

6.4. ИНТЕРПОЛЯЦИОННАЯ ФОРМУЛА НЬЮТОНА

Значения некоторого многочлена оцениваются в точках a, $a + h$, $a + 2h$, . . ., $a + nh$. На основе этих значений формируется таблица разностей. В интерполяционной формуле Ньютона для разностей вперед утверждается, что значение рассматриваемого многочлена в точке $a + rh$ равно:

$$f(a + rh) = f(a) + \binom{r}{1}\Delta f(a) + \binom{r}{2}\Delta^2 f(a) + \ldots \qquad (6.4.1)$$

Эта формула представляет собой конечно-разностный аналог разложения в ряд Тейлора (1.2.1) и она точна для всех вещественных значений r, если $f(x)$ — многочлен. Результаты, полученные с ее помощью, точные и для других функций, если r — целое положительное число, которое меньше n или равно ему. Эта формула вряд ли представляла бы какой-либо интерес, если бы ее применение ограничивалось подобными случаями. К счастью, данную формулу можно применять и в других ситуациях. (Это демонстрируется далее во втором примере.) Получаемые результаты являются приближенными, однако приближение может быть очень точным.

Следует обратить внимание на тот факт, что дифференциальные операторы и биномиальные коэффициенты [2] в формуле (6.4.1) можно рассматривать как результат разложения по формуле бинома оператора $(1 + \Delta)^r$. Поэтому формулу Ньютона можно записать в виде символической операторной формулы

$$f(a + r\Delta) = (1 + \Delta)^r f(a). \qquad (6.4.2)$$

Пример 6.4.1. С помощью табл. 6.2.1 найдем значения многочлена $f(x) = 8x^3 - 2x + 1$ в точке $x = 0{,}25$.

В формулу (6.4.1) подставляем $a = 0$, $h = 0{,}5$ и $r = 0{,}5$. В соответствии с этой формулой получаем

$$f(0{,}25) = f(0) + 0{,}5\,\Delta f(0) - 0{,}125\,\Delta^2 f(0) + 0{,}0625\,\Delta^3 f(0) =$$
$$= 1 + 0{,}5 \times 0 - 0{,}125 \times 6 + 0{,}0625 \times 6 = 0{,}625.$$

[2] См. параграф 1.5.

Можно подставить и другие значения: $a = 1$, $h = 0{,}5$ и $r = -1{,}5$. Формула (6.4.1) в этом случае дает

$$f(0{,}25) = f(1) - 1{,}5\,\Delta f(1) + 1{,}875\,\Delta^2 f(1) - 2{,}1875\,\Delta^3 f(1) =$$
$$= 7 - 1{,}5 \times 18 + 1{,}875 \times 18 - 2{,}1875 \times 6 = 0{,}625.$$

В правильности полученного ответа можно убедиться непосредственным вычислением.

Пример 6.4.2. В табл. 6.4.1 приведены значения площади под стандартной нормальной кривой слева от точки x для некоторых значений x, взятых с интервалом 0,1. Найдем площадь под этой кривой слева от точки $x = 1{,}96$.

Вычисления начинаем с составления таблицы разностей, приведенной далее. Значения разностей третьего порядка малы, поэтому мы не будем принимать во внимание разности четвертого и более высоких порядков. Площадь слева от точки $x = 1{,}96$ можно вычислить по формуле (6.4.1) при $a = 1{,}9$, $h = 0{,}1$ и $r = 0{,}6$:

$$f(1{,}96) = f(1{,}9) + 0{,}6\,\Delta f(1{,}9) - 0{,}12\,\Delta^2 f(1{,}9) + 0{,}056\,\Delta^3 f(1{,}9) =$$
$$= 0{,}9713 + 0{,}6 \times 0{,}0060 - 0{,}12 \times (-0{,}0012) +$$
$$+ 0{,}056 \times 0{,}0004 = 0{,}9750.$$

Сравнивая полученный результат с соответствующим значением в таблице нормального распределения, приведенной на с. 327, обнаруживаем, что ответ найден правильно с точностью до четырех знаков после запятой.

Литература: [27, с. 29—35, 81—90], [41, с. 60—62, 64, 92—94], [43, с. 94—95], [84, с. 52], [92, с. 550—552].

Таблица **6.4.1. Площадь под стандартной нормальной кривой слева от точки x**

x	Площадь $f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$
1,9	0,9713			
		0,0060		
2,0	0,9773		−0,0012	
		0,0048		0,0004
2,1	0,9821		-0,0008	
		0,0040		0,0000
2,2	0,9861		-0,0008	
		0,0032		
2,3	0,9893			

6.5. ИНТЕРПОЛЯЦИОННАЯ ФОРМУЛА БЕССЕЛЯ

Интерполяционная формула Ньютона является точной лишь тогда, когда рассматриваемая функция представляет собой многочлен. В других случаях данный метод можно применять лишь

как приближенный и получаемое при этом приближение может не быть столь точным, как это необходимо. Например, для получения значения функции в точке $x = a + \frac{1}{2}$ A используются табличные значения, соответствующие точкам a, $a + A$, $a + 2A$, $a + 3h$ ... Все эти точки (кроме первой) лежат справа от рассматриваемой точки интерполяции. Более удовлетворительный ответ получают, как правило, при применении формулы, более или менее симметричной относительно точки интерполяции. Одна из таких формул — интерполяционная формула Бесселя, она обеспечивает точный ответ для случаев, когда значение r лежит в интервале $(0,1)$:

$$f(a + rh) = f(a) + r\,\Delta f(a) + \frac{1}{4}(r^2 - r)(\Delta^2 f(a) + \Delta^2 f(a - h)) + \frac{1}{2}(r^2 - r)\left(r - \frac{1}{2}\right)\Delta^3 f(a - h) + \frac{1}{48}(r^3 - r)(r - 2)(\Delta^4 f(a - 2h)) + \ldots \qquad (6.5.1)$$

Пример 6.5.1. Заданы следующие значения функции:

x	$f(x)$	x	$f(x)$
0,7	1,428 571 4	1,1	0,909 090 9
0,8	1,250 000 0	1,2	0,833 333 3
0,9	1,111 111 1	1,3	0,769 230 8
1,0	1,000 000 0		

Вычислим $f(0,95)$.

Воспользуемся интерполяционной формулой Ньютона (6.4.1) и таблицей разностей 6.5.1. Выбираем $a = 0,9$, $h = 0,1$ и $r = 0,5$ и получаем значение 1,052 654 1.

Таблица 6.5.1. Таблица разностей к примеру 6.5.1

$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$
1,428 571 4				
	−0,178 571 4			
1,250 000 0		0,039 682 5		
	−0,138 888 9		−0,011 904 7	
1,111 111 1		0,027 777 8		0,004 328 9
	-0,111 111 1		−0,007 575 8	
1,000 000 0		0,020 202 0		0,002 525 3
	−0,090 909 1		−0,005 050 5	
0,909 090 9		0,015 151 5		0,001 554 1
	−0,075 757 6		−0,003 496 4	
0,833 333 3		0,011 655 1		
	−0,064 102 5			
0,769 230 8				

По формуле Бесселя при $a = 0,9$, $h = 0,1$ и $r = 0,5$ получаем значение 1,0526372.

Этот пример выбран в иллюстративных целях и было известно, что $f(x) = 1/x$. Следовательно, истинное значение $f(0,95)$ равно 1,052 631 6. Мы видим, что обе формулы дали достаточно точные результаты, но ответ, полученный по формуле Бесселя, точнее, чем по формуле Ньютона.

Литература: [27, с. 61—72], [41, с. 67—72], [42, с. 222—227], [43, с. 97—110].

6.6. ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

Ситуации, приводящие к численному дифференцированию, описаны в параграфе 6.1. Важно помнить, что при применении численного метода дифференцирования на самом деле производится оценка тангенса угла наклона аппроксимирующей полиномиальной дуги. Эта дуга проходит ниже и выше соответствующих значений функции. Отсюда следует, что численное дифференцирование — весьма рискованная процедура, и по возможности его нужно избегать.

Предположим, что функция $f(x)$ может быть представлена удовлетворительным образом с помощью некоторого многочлена второй или меньшей степени в окрестности точки $x = a$ и что известны значения $f(a - A)$, $f(a)$ и $f(a + h)$. Обозначим эти значения символами f_{-1}, f_0 и f_1 соответственно. Тогда

$$f'_{-1} = \frac{1}{2h}(-3f_{-1} + 4f_0 - f_1), \qquad (6.6.1)$$

$$f'_0 = \frac{1}{2h}(-f_{-1} + f_1), \qquad (6.6.2)$$

$$f'_1 = \frac{1}{2h}(f_{-1} - 4f_0 + 3f_1). \qquad (6.6.3)$$

Формулы (6.6.1) и (6.6.3) необходимо использовать лишь в начале и конце таблицы.

Если функция $f(x)$ может быть представлена удовлетворительным образом с помощью многочлена четвертой или меньшей степени, применяют следующие формулы:

$$f'_{-2} = \frac{1}{12h}(-25f_{-2} + 48f_{-1} - 36f_0 + 16f_1 - 3f_2), \qquad (6.6.4)$$

$$f'_{-1} = \frac{1}{12h}(-3f_{-2} - 10f_{-1} + 18f_0 - 6f_1 + f_2), \qquad (6.6.5)$$

$$f'_0 = \frac{1}{12h}(f_{-2} - 8f_{-1} + 8f_1 - f_2), \qquad (6.6.6)$$

$$f'_1 = \frac{1}{12h}(-f_{-2} + 6f_{-1} - 18f_0 + 10f_1 + 3f_2), \qquad (6.6.7)$$

$$f'_2 = \frac{1}{12h}(3f_{-2} - 16f_{-1} + 36f_0 - 48f_1 + 25f_2). \qquad (6.6.8)$$

Все эти формулы легко обосновать, раскладывая в ряд Тейлора выражения, стоящие в правой части формул, относительно точки, которой соответствует индекс при производной в левой части. Формулы (6.6.4), (6.6.5), (6.6.7) и (6.6.8) следует использовать лишь в начале и конце таблицы.

Необходимо подчеркнуть, что все эти формулы можно применять лишь к достаточно «гладким» данным. При наличии сомнений относительно подходящего аппроксимирующего многочлена формула для пяти точек более пригодна, чем формула для трех точек.

Вычисление второй производной еще более рискованно, чем вычисление первой. Можно дважды воспользоваться формулами, приведенными выше, но удобнее, как правило, одна из следующих:

$$f_0'' = \frac{1}{h^2}(f_{-1} - 2f_0 + f_1), \qquad (6.6.9)$$

$$f_0'' = \frac{1}{12h^2}(-f_{-2} + 16f_{-1} - 30f_0 + 16f_1 - f_2). \qquad (6.6.10)$$

При применении формулы для трех точек предполагают, что аппроксимирующий многочлен имеет степень два, пяти точкам соответствует многочлен четвертой степени.

Для численного дифференцирования применимы также и методы конечных разностей, хотя методы, описанные выше, более предпочтительны. Отправным моментом рассуждения в данном случае является ряд Тейлора (1.2.1). Будем писать D вместо d/dx, D^2 вместо d^2/dx^2 и т. д. Тогда

$$f(a+rh) = f(a) + \frac{rh}{1!}Df(a) + \frac{(rh)^2}{2!}D^2f(a) + \ldots =$$

$$= \left(1 + \frac{rhD}{1!} + \frac{(rhD)^2}{2!} + \ldots\right)f(a) = e^{rhD}f(a).$$

Сопоставим эту формулу с (6.4.2). Видим, что оператор e^{hD} эквивалентен оператору $1 + \Delta$ и что

$$Df(a) = \left(\frac{1}{h}\ln(1+\Delta)\right)f(a) =$$

$$= \frac{1}{h}\left(\Delta f(a) - \frac{1}{2}\Delta^2 f(a) + \frac{1}{3}\Delta^3 f(a) - \ldots\right), \qquad (6.6.11)$$

$$D^2f(a) = \left(\frac{1}{h}\ln(1+\Delta)\right)^2 f(a) =$$

$$= \frac{1}{h^2}\left(\Delta^2 f(a) - \Delta^3 f(a) + \frac{11}{12}\Delta^4 f(a) + \ldots\right). \qquad (6.6.12)$$

Эти формулы точны для многочленов и приближенны для прочих функций. Их использование связано с теми же трудностями, что и применение интерполяционной формулы Ньютона (см. параграф 6.5). Этот метод иногда бывает полезен в тех случаях, когда необходимо вычислить значение производной в нетабличной точке (см. пример 6.6.3).

Пример 6.6.1. В табл. 6.6.1 даны значения функции $\operatorname{arctg}(e^x)$. Вычислим первую и вторую производные этой функции в точке $x = 1{,}2$.

Формулы, задающие явное аналитическое выражение для этих двух производных, довольно сложны. Воспользуемся формулой (6.6.6) для вычисления значения первой производной:

$$f'(1{,}2) = \frac{1}{12\times 0{,}1}(1{,}218283 - 8\times 1{,}249\,462 +$$

$$+ 8\times 1{,}304\,726 - 1{,}329\,023) = 0{,}276\,143.$$

В этом ответе точны все шесть цифр.

Значение второй производной (тангенс угла наклона на графике) можно вычислить по формуле (6.6.10):

$$\Gamma(1.2) = \frac{1}{12\times(0{,}1)^2}(-1{,}218\,283 + 16\times 1{,}249\,462 -$$

$$- 30\times 1{,}278244 + 16\times 1{,}304\,726 - 1{,}329\,023 = -0{,}230\,15.$$

Истинное значение равно —0,230 21; итак, полученный ответ точен до четырех знаков после запятой.

Таблица 6.6.1. **Значение функции $\operatorname{arctg}(e^x)$**

x	$\operatorname{arctg}(e^x)$	x	$\operatorname{arctg}(e^x)$
1,0	1,218283	1,3	1,304 726
1,1	1,249 462	1,4	1,329 023
1,2	1,278 244		

Пример 6.6.2. Найдем тангенс угла наклона в точке $x = 60$ графика функции, описывающей некоторую эмпирическую зависимость; значения этой функции приведены в табл. 6.6.2.

По формуле (6.6.4) получаем значение тангенса угла наклона — 1491. Можно воспользоваться также методом конечных разностей. Разности третьего порядка мало различаются и относительно малы по абсолютной величине:

$$f'(60) = A/(60) - \frac{1}{2}\Delta^2 f(60) + \frac{1}{3}\Delta^3 f(60) =$$

$$= -1536 - \frac{1}{2}(-97) + \frac{1}{6}(-8) = -1490.$$

Пример 6.6.3. Найдем тангенс угла наклона в точке $x = 61{,}5$ графика функции, значения которой даны в табл. 6.6.2.

Таблица 6.6.2. **Таблица разностей к примеру 6.6.2**

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
60	63 620			
		− 1536		
61	62 084		−97	
		−1633		--8
62	60 451		-105	
		-1738		−6
63	58 713		-111	
		−1849		−5
64	56 864		-116	
		− 1935		0
65	54 899		-116	
		−2081		−1
66	52 818		−117	
		- 2198		
67	50 620			

По формулам (6.6.11) и (6.4.2) получаем:

$$f'(61{,}5) = (\ln(1+\Delta))\, f(61{,}5) = (\ln(1+\Delta))(1+\Delta)^{\frac{1}{2}} f(61) =$$

$$= \left(\Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \ldots\right)\left(1 + \frac{1}{2}\Delta - \right.$$

$$\left. - \frac{1}{8}\Delta^2 + \ldots\right) f(61) = \left(\Delta - \frac{1}{24}\Delta^3 + \ldots\right) f(61) =$$

$$= \Delta f(61) - \frac{1}{24}\Delta^3 f(61) = -1633 - \frac{1}{24}(-6) = -1633.$$

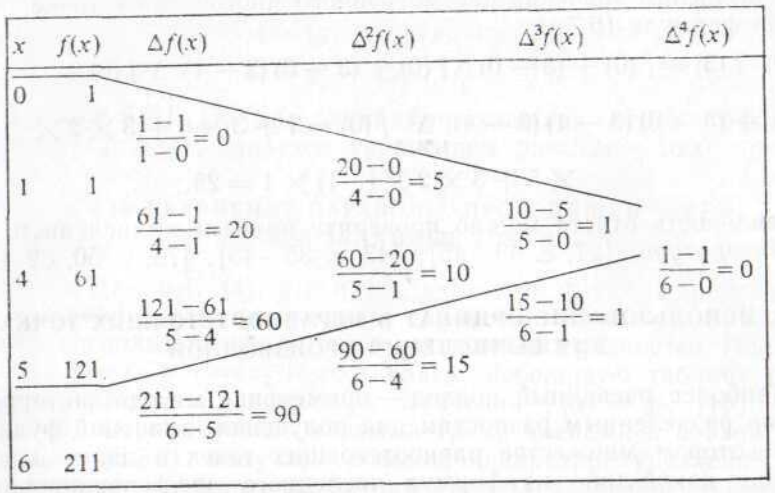
Литература: *[27,* с. 126—127], [43, с. 82, 134], [73, с. 56], [84, с. 83—85].

6.7. ОРДИНАТЫ В НЕРАВНООТСТОЯЩИХ ТОЧКАХ — ИНТЕРПОЛЯЦИЯ ПО РАЗДЕЛЕННЫМ РАЗНОСТЯМ

До сих пор мы рассматривали конечно-разностные формулы, в которых предполагалось, что используются значения функций на некотором множестве равноотстоящих точек. Однако бывают и другие случаи, в связи с этим необходимо ввести понятие *разделенных разностей.*

Построение таблицы разделенных разностей для многочлена третьей степени $f(x) = x^3 - x + 1$ показано в табл. 6.7.1. Мы приводим систему обозначений Фримена. Так, разделенная разность третьего порядка для $f(a)$, при подсчете которой берутся значения $f(a)$, $f(b)$, $f(c)$ и $f(d)$, обозначается символом $\Delta^3 f(a)$. При выполнении конкретных вычислений для заполнения таблицы следует очень тщательно оформлять записи. Например, $\underset{1,4,5}{\Delta^3 f}(0)$ вычисляется следующим образом:

Таблица 6.7.1. **Разделенные разности**



Примечание. Проведенные наклонные линии показывают процесс вычисления $\underset{1,4,5}{\Delta^3} f(0)$

$$\underset{1.4,5}{\text{дз}}\, f(0) = \left(\underset{4,5}{\Delta^2} f(1) - \underset{1,4}{\Delta^2} f(0)\right)/(5-0) = (10-5)/(5-0) = 1.$$

Это число помещается в столбце разностей третьего порядка в той клетке таблицы, которая соответствует вершине равнобедренного треугольника с основанием, совпадающим с клетками, где записаны значения ДО), $f(1)$, $f(4)$ и $f(5)$. Следует заметить, что $f(x)$ представляет собой многочлен третьей степени, поэтому разделенные разности третьего порядка равны между собой.

Интерполяционная формула Ньютона для разделенных разностей имеет вид

$$f(x) = f(a) + (x - a)\,\underset{b}{\Delta f}(a) + (x-a)(x-b)\,\underset{b,c}{\Delta^2} f(a) + \ldots \quad (6.7.1)$$

Необходимо обратить внимание на тот факт, что для функции, являющейся многочленом, значения функции могут располагаться в соответствующем столбце таблицы разделенных разностей в произвольном порядке, и получаемый при этом результат всегда будет точен. Для функций, не являющихся многочленами, вычисления по формуле Ньютона обычно дают наилучшие результаты, если значения функции расположены в естественном порядке и аргументы используемых разделенных разностей близки, насколько это возможно, к рассматриваемому значению аргумента. Интерполяционная формула Ньютона для разностей вперед (6.4.1) — особый случай формулы (6.7.1).

Пример 6.7.1. С помощью таблицы разделенных разностей 6.7.1 вычислим значения рассмотренного многочлена в точке $x = 3$. По формуле (6.7.1)

$$f(3) = f(0) + (3-0)\underset{1}{\Delta} f(0) + (3-0)(3-1)\underset{1,4}{\Delta^2} f(0) +$$

$$+ (3-0)(3-1)(3-4)\underset{1,4,5}{\Delta^2} f(0) = 1 + 3\times 0 + 3\times 2\times$$

$$\times 5 + 3\times 2\times(-1)\times 1 = 25.$$

Правильность ответа можно проверить прямым вычислением.

Литература: [27, с. 39—45], [43, с. 35—45], [73, с. 50, 82—87].

6.8. ИСПОЛЬЗОВАНИЕ ОРДИНАТ В НЕРАВНООТСТОЯЩИХ ТОЧКАХ ДЛЯ ВЫЧИСЛЕНИЯ ПРОИЗВОДНОЙ

Наиболее очевидный подход — применение метода интерполяции по разделенным разностям для получения значений функций на некотором множестве равноотстоящих точек и далее использование какой-либо из формул численного дифференцирования, приведенных в параграфе 6.6. Для получения значений в равноотстоящих точках можно также воспользоваться методом сплайна, описанным в параграфе 4.3. В любом случае необходимо выполнить большой объем вычислений. Наилучший же подход состоит в том, чтобы продифференцировать соотношение (6.7.1):

$$f'(x) = \underset{b}{\Delta} f(a) + ((x-a)+(x-b))\underset{b,c}{tf} f(a) + ((x-a)(x-b)+$$

$$+ (x-a)(x-c)+(x-b)(x-c))\underset{b,c,d}{\Delta^3} f(a) + \ldots \qquad (6.8.1)$$

Эта формула дает точные результаты для многочленов. Для функций, не являющихся многочленами, результаты будут приближенными, при этом не следует забывать об опасности возникновения крупной ошибки при численном дифференцировании (см. параграф 6.6).

Пример 6.8.1. Найдем производную многочлена, значения которого приведены в табл. 6.7.1, в точке $x = 2$.

Воспользуемся формулой (6.8.1) при $a = 0$, $b = 1$, $c = 4$ и $d = 5$:

$$f'(2) = 0 + ((2-0)+(2-1))\times 5 + ((2-0)(2-1)+$$

$$+ (2-0)(2-4)+(2-1)(2-4))\times 1 = 11.$$

Этот результат можно проверить непосредственным вычислением.

6.9. УРАВНЕНИЕ ПРЯМОЙ, ПРОХОДЯЩЕЙ ЧЕРЕЗ ДВЕ ЗАДАННЫЕ ТОЧКИ

Пусть (x_0, y_0) и (x_1, y_1) — Две различные точки. Уравнение прямой линии, проходящей через эти две точки, имеет вид

$$y = Ax + B, \qquad (6.9.1)$$

где

$$\left.\begin{aligned} A &= (y_1 - y_0)/(x_1 - x_0) \\ B &= y_0 - Ax_0 \end{aligned}\right\}. \qquad (6.9.2)$$

Пример 6.9.1. Прямая линия, проходящая через точки (37; 324) и (42; 584), задается уравнением $y = 52x - 1600$.

6.10. УРАВНЕНИЕ ПАРАБОЛЫ, ПРОХОДЯЩЕЙ ЧЕРЕЗ ТРИ ЗАДАННЫЕ ТОЧКИ

Пусть (x_0, y_0), (x_1, y_1) и (x_2, y_2) — три точки с различными абсциссами[3]. Для проведения параболы через эти три точки можно воспользоваться методом разделенных разностей. Полагаем $a = x_0$, $b = x_1$ и $c = x_2$ и составляем небольшую таблицу разделенных разностей, используя значения $f(a) = y_0$, $f(b) = y_1$ и $f(c) = y_2$. Ограничиваясь первыми тремя членами в правой части формулы (6.7.1), получаем искомое квадратное уравнение. Этот метод можно обобщить на случай многочлена любой степени.

Далее предлагается альтернативный метод, который не требует предварительного знакомства с понятием разделенных разностей и непосредственно приводит к искомому уравнению. Вычисляем значения следующих величин:

$$\left.\begin{aligned} a &= x_1 - x_0, \\ b &= x_2 - x_0, \\ \alpha &= y_1 - y_0, \\ \beta &= y_2 - y_0, \\ A &= -(a\beta - b\alpha)/\{ab(a-b)\}, \\ B &= (a^2\beta - b^2\alpha)/\{ab(a-b)\}. \end{aligned}\right\} \qquad (6.10.1)$$

J

Квадратное уравнение имеет вид

$$y = A(x - x_0)^2 + B(x - x_0) + y_0. \qquad (6.10.2)$$

Обратите внимание на тот факт, что рассматриваемые три точки могут быть взяты в произвольном порядке.

Пример 6.10.1. Найдем уравнение параболы, проходящей через точки (0; 1), (0,5; 0,7788) и (2; 0,3679). Вычисляем:

$$a = 0{,}5; \qquad \alpha = -0{,}2212;$$

$$b = 2{,}0; \qquad \beta = -0{,}6321;$$

$$A = 0{,}08423\dot{}; \qquad B = -0{,}484516\dot{}.$$

Итак, искомое квадратное уравнение имеет вид

$$y = 0{,}08423\dot{}x^2 - 0{,}484516\dot{}x + 1{,}0.$$

[3] С различными значениями переменной x.

6.11. УПРАЖНЕНИЯ

1. а) Составьте таблицу разностей для многочлена $f(x)=x^4$, взяв значения этой функции в точках $x=1, 2, 3, \ldots, 8$.

б) На основании того факта, что все разности четвертого порядка в данной таблице равны между собой, заполните ее дополнительную строку и получите значение $f(9)$.

в) С помощью интерполяционной формулы Ньютона для разностей вперед получите значение $f(1,5)$.

г) Воспользуйтесь одним из численных методов дифференцирования для вычисления значения $f'(1)$.

д) Воспользуйтесь одним из численных методов для вычисления значения $f''(5)$.

е) Примените метод из примера 6.6.3 для оценки значения $f'(1,5)$.

Проверьте все полученные ответы непосредственным вычислением.

2. Функция пяти переменных x_1, x_2, x_3, x_4 и x_5 задана уравнением

$$y=\sum_{j=2}^{4}\frac{(x_{j+1}-2x_j+x_{j-1})^2}{(x_{j+1}-x_j)^2+(x_j-x_{j-1})^2}.$$

Вычислите значение частной производной y по x_3 в точке $x_1=1$, $x_2=4$, $x_3=10$, $x_4=17$, $x_5=24$.

3. С помощью формулы (6.7.1) и значений многочлена в точках $x=1, 4, 5$ и 6, приведенных в табл. 6.7.1, вычислите значения $f(3)$.

7. НЕКОТОРЫЕ ДРУГИЕ ЧИСЛЕННЫЕ МЕТОДЫ

В этой главе описано много полезных численных методов. Мы начнем с задачи перегруппировки данных, затем перейдем к формуле Харди, которая позволяет оценить центральную ординату площади. Следующая важная тема — метод согласования двух гладких кривых, обеспечивающий проведение единой кривой. В параграфе 7.5 описан метод наискорейшего спуска для минимизации функции нескольких переменных, который находит применение во многих областях. В частности, на нем основан нелинейный метод наименьших квадратов (см. гл. 18). Глава завершается описанием простого приема, позволяющего увеличить емкость зоны хранения данных в программируемом калькуляторе.

7.1. ПЕРЕГРУППИРОВКА СГРУППИРОВАННЫХ ДАННЫХ

Данные часто бывают сгруппированными. Это делается либо для того, чтобы получить краткую таблицу, либо потому, что значения наблюдений очень разбросаны, либо по какой-нибудь другой причине. Исследователь может найти, что группировка данных в таблице, с которой он работает, ему не подходит и он должен их перегруппировать. Это может оказаться очень трудной задачей, особенно если число случаев в отдельных группах значительно превышает число случаев в соседних группах. Наилучший подход к решению этой задачи — применение интерполяционного метода к *кумулятивным суммам,*

Пример 7.1.1, В 1968 г. в Австралии число детей, рожденных матерями, не состоящими в браке, равнялось 18980. Совокупность

этих матерей представлена в табл. 7.1.1 большими возрастными группами. Оценим число детей, рожденных незамужними матерями, по каждой из следующих пятилетних возрастных групп: 12—16, 17—21, 22—26, . . ., 47—51.

Таблица 7.1.1. **Число детей, рожденных матерями, не состоящими в браке; Австралия, 1968 г.**

Число полных лет матери	Число родившихся детей
14 и менее	121
15-19	7 169
20-29	8 857
30-39	2 427
40-49	406
50 и более	0
	Всего 18 980

Начинаем решение задачи с вычисления накопленных итогов. Число детей, рожденных незамужними матерями, которые не достигли

15 лет — 121,
20 лет — 7 290,
30 лет — 16 147,
40 лет — 18574,
50 лет — 18 980.

Теперь подберем гладкую кривую, проходящую через точки кумулятивных итогов (15, 121), (20, 7290), (30, 16 147), (40, 18574) и (50, 18980). Можно применить любой подходящий интерполяционный метод; в подобной ситуации чаще всего пользуются графическим методом. Мы возьмем сплайн-функцию (см. параграф 4.3) и будем интерполировать назад от точки 50.

Число детей, рожденных незамужними матерями в возрасте около 50 лет, очень невелико; таким образом, начальный многочлен можно выбрать так, чтобы его кривая проходила через точки (50, 18980) и (40, 18574), причем первая и вторая производные в точке 50 равны нулю. Из построенной сплайн-кривой получим следующие оценки:

Число детей, рожденных незамужними матерями в возрасте

до 17 лет — 1 782,
до 22 лет — 10 415,
до 27 лет — 14843,
до 32 лет — 16 825,
до 37 лет — 18 099,
до 42 лет — 18772,
до 47 лет — 18969,
до 50 лет — 18 980.

Производя соответствующие вычисления, можно получить результаты, приведенные в табл. 7.1.2 (принято предположение, что

число детей, рожденных матерями в возрасте до 12 лет, равно нулю).

Таблица 7.1.2. **Число детей, рожденных матерями, не состоящими в браке; Австралия, 1968 г. (числовые оценки по пятилетним возрастным группам)**

Число полных лет матери	Число родившихся детей
12-16	1 782
17-21	8 633
22-26	4 428
27-31	1 982
32-36	1 274
37-41	673
42-46	197
47-51	11
	Всего 18980

7.2. ЦЕНТРАЛЬНАЯ ОРДИНАТА ПЛОЩАДИ, ФОРМУЛА ХАРДИ

Иногда возникает ситуация, когда известны площади участков под кривой, расположенных между последовательными равноотстоящими ординатами, и необходимо определить ординату кривой посередине между ними. Ответ можно получить с помощью формулы Харди, применение которой рассматривается как специальный метод численного дифференцирования.

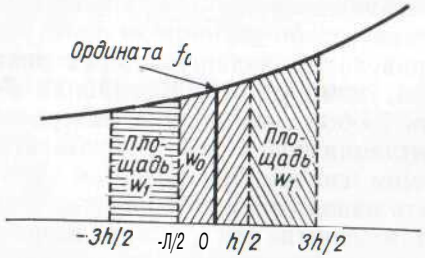


Рис. 7.2.1. Применение формулы Харди для оценки ординаты кривой

Рассмотрим кривую, описываемую неизвестной функцией $f(x)$ в интервале от $(-3h/2)$ до $3h/2$. Про функцию $f(x)$ известно, что площади под кривой, заключенные между вертикалями, которые исходят из точек $x = -3h/2$, $x = -h/2$, $x = h/2$ и $x = 3h/2$, равны соответственно w_{-1}, w_0, w_1. Эти площади обозначены на рис. 7.2.1. Приближенно ордината кривой в точке $x = 0$ дается формулой [1]

$$f(0) = \frac{1}{h}\left(w_0 - \frac{1}{24}\Delta^2 w_{-1}\right). \qquad (7.2.1)$$

Эта формула является точной, когда $f(x)$ представляет собой многочлен третьей степени.

[1] Оператор конечных разностей Д определен в параграфе 6.2. Здесь, однако, достаточно заметить, что $\Delta^2 w_{-1} = w_1 - 2w_0 + w_{-1}$.

Пример 7.2.1. В табл. 7.2.1 показаны в различные моменты времени значения среднего веса в группах самцов крыс, получивших различные дозы монотретичного бутилгидрохинона. Вычислим дневной прирост веса на 135-й день для каждой дозы.

Вес крысы в момент времени x равен интегралу дневного прироста за период от зачатия до момента x. Мы должны вычислить дневной прирост веса на 135-й день, который лежит посередине между 128-м и 142-м днем. Для этого можно воспользоваться формулой Харди (7.2.1).

Площадь под кривой дневного прироста веса между вертикалями в точках $x = 128$ (дней) и $x = 142$ (дня) равна изменению веса за этот период. Следовательно, для контрольной группы

$$w_{-1} = 586 - 558 = 28,$$
$$w_0 = 599 - 586 = 13,$$
$$w_1 = 601 - 599 = 2.$$

Дневной прирост веса в средней точке (135-й день) равен:

$$f(0) = \frac{1}{14}\left(13 - \frac{1}{24}(2 - 2 \times 13 + 28)\right) = 0{,}917 \text{ г/дн.}$$

Аналогичные значения, полученные для 0,02 %-ной, 0,1 %-ной и 0,5 %-ной доз монотретичного бутилгидрохинона, соответственно равны 1,065 г/дн., 1,378 г/дн. и 1,366 г/дн.

Таблица 7.2.1. **Влияние монотретичного бутилгидрохинона на вес самцов крыс**

Дни эксперимента	Средний вес, г			
	Доза, %			Контрольная группа
	0,02	0,1	0.5	
86	535	504	475	503
100	568	529	508	535
114	575	545	528	558
128	614	588	548	586
142	630	608	567	599
156	649	622	582	601
170	667	627	595	616

Источник. Личное сообщение К. Д. Кейрнкросса (Школа биологических наук, Университет Маккуори, Норс Рид 2113, Австралия).

Литература: [96, с. 21—22].

7.3. ЦЕНТРАЛЬНАЯ ОРДИНАТА СУММЫ ОРДИНАТ

Формула Харди (7.2.1) позволяет вычислить ординату кривой по значениям площади участков под ней. Можно вывести аналогичную формулу для средней ординаты группы ординат, сложенных вместе.

Рассмотрим неизвестную функцию $f(x)$. Выразим значение функции $f(x)$ в точке $x = a + rh$ через f_r $(r = \ldots, -2, -1, 0, 1, 2, \ldots)$. Про функцию $f(x)$ известно только то, что сумма n центральных ординат равна w_0, сумма предшествующих n ординат равна w_{-1}, а сумма последующих n ординат равна w_1. Целое число n нечетное.

Приближенное значение средней ординаты определяется формулой [2]

$$f_0 = \frac{1}{n}\left(w_0 - \frac{n^2-1}{24n^2}\Delta^2 w_{-1}\right). \qquad (7.3.1)$$

Эта формула является точной, когда $f(x)$ представляет собой многочлен третьей степени.

Пример 7.3.1. Сглаживающий эффект скользящего среднего был отмечен в параграфе 4.2. Простой метод скользящего среднего, однако, имеет тенденцию к искажению значений гладкой функции. Здесь возникает проблема оценки искажения. Сумма n последовательных наблюдений обеспечит неискаженную оценку суммы n значений рассматриваемой гладкой функции, и формула (7.3.1) может тогда быть использована для оценки центральной ординаты.

Воспользуемся формулой (7.3.1) при $n = 5$, чтобы оценить «сглаженные» ординаты в точках $x = 42$, 52, 62 для данных из табл. 4.2.1.

В точке $x = 42$

$$w_{-1} = 191 + 419 + 278 = 1621,$$

$$w_0 = 381 + 384 + 665 + 477 + 1015 = 2922,$$

$$w_1 = 1093 + 860 + 779 + 862 + 951 = 4545$$

и

$$\Delta^2 w_{-1} = 4545 - 2 \times 2922 + 1621 = 322.$$

Сглаженное значение в точке $x = 42$ равно, следовательно,

$$\frac{1}{5}\left\{2922 - \frac{25-1}{24\times 25}(322)\right\} = 581{,}82.$$

Аналогично находятся сглаженные значения в точках $x = 52$ и $x = 62$; они равны соответственно 1203,78 и 2648,93.

Метод, приведенный в параграфе 6.10, может применяться для подбора параболы, проходящей через указанные три точки, полученная кривая показана на рис. 7.3.1. В области изменения x от $x = 42$ до $x = 62$ такой подбор имеет смысл, но вне данной области это не так.

[2] См. сноску 1 в параграфе 7.2

Формула (7.3.1) может быть использована для получения сглаженных значений $x = 47$ и $x = 57$, а сплайн-метод (см. параграф 4.3) может тогда применяться для соединения всех пяти

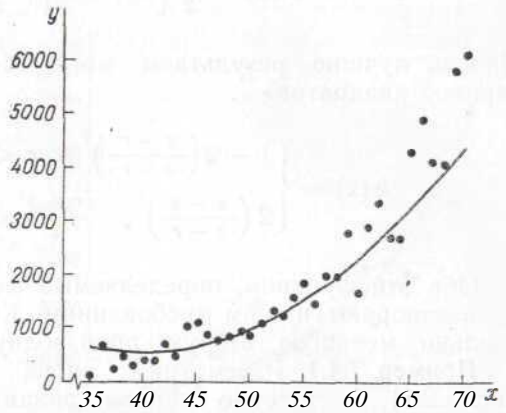


Рис. 7.3.1. Парабола, построенная по 36 точкам

точек. Однако через сглаженные значения $x = 47$ и $x = 57$ единую гладкую кривую провести особенно трудно.

Литература: [96, с. 20—22], [100, с. 57—59; русский перевод с. 56—58].

7.4. СОГЛАСОВАНИЕ ДВУХ ГЛАДКИХ КРИВЫХ

Предположим, что подобрана кривая к данным наблюдений в области изменения x от $x = 0$ до $x = s$, а вторая кривая подбирается к данным наблюдений в области изменения x от $x = r$ п далее $(r < s)$. Две кривые накладываются друг на друга в области изменения x, определяемой неравенством $r \leqslant x \leqslant s$. Задача состоит в том, чтобы согласовать значения функций, соответствующие общей области изменения x, таким образом, чтобы график полученной в итоге функции плавно переходил с первой кривой па вторую.

Обозначим функцию, соответствующую первой кривой, через $g(x)$, функцию, соответствующую второй кривой,— через $h(x)$ п «смешивающую функцию» — через $k(x)$. В области изменений значений x, определяемых неравенством $r \leqslant x \leqslant s$, мы ищем функцию

$$f(x) = k(x)\,g(x) + \{1 - k(x)\}\,h(x). \qquad (7.4.1)$$

Для плавного перехода нужно, чтобы выполнялись следующие соотношения:

$$k(r) = 1, \quad k'(r) = 0, \quad k''(r) \text{ мало};$$

$$k(s) = 0, \quad k'(s) = 0, \quad k''(s) \text{ мало}.$$

Обычно, хотя это не так существенно, смешивающая кривая выбирается симметричной относительно точки $(r+s)/2$. Так называе-

мая «кривая синусов» часто используется как смешивающая функция

$$k(x) = \frac{1}{2}\left[1 + \cos\left\{\frac{(x-r)\pi}{s-r}\right\}\right]. \qquad (7.4.2)$$

Иногда лучшие результаты могут быть получены с помощью «кривой квадратов»

$$k(x) = \begin{cases} 1 - 2\left(\frac{x-r}{s-r}\right)^2, & r < x < \frac{r+s}{2}; \\ 2\left(\frac{x-s}{s-r}\right)^2, & \frac{r+s}{2} < x < s. \end{cases} \qquad (7.4.3)$$

Обе эти функции, определяемые формулами (7.4.2) и (7.4.3), удовлетворяют нашим требованиям. Кривая квадратов имеет несколько меньшую вторую производную в точках $x = r$ и $x = s$.

Пример 7.4.1. Рассмотрим снова данные табл. 4.2.1 (и рис. 4.2.1). Достаточно гладкая кривая может быть получена в результате подбора параболы, проходящей через точки (40, 381), (47, 779) и (58, 2011), параболы, проходящей через точки (50, 866), (58, 2011) и (69, 5835), и смешивания этих двух кривых с помощью кривой синусов.

Параболы, представленные в табл. 7.4.1, были получены методом, описанным в параграфе 6.10. В этой таблице приведена также функция, полученная в результате смешивания. На рис. 7.4.1 можно увидеть, что переход от первой параболы ко второй не так гладок, как бы нам этого хотелось. Изменяя центральные точки

Таблица 7.4.1. **Согласование двух парабол с помощью кривой синусов**

x	Первая парабола	Вторая парабола	Получен-ная кривая	x	Первая парабола	Вторая парабола	Получен-ная кривая
35	281	—	281	53	1359	1134	1281
36	288		288	54	1477	1266	1372
37	302	—	302	55	1601	1420	1483
38	322	—	322	56	1732	1596	1624
39	349		349	57	1868	1793	1800
40	381	—	381	58	2011	2011	2011
41	419	—	419	59	—	2251	2251
42	464	—	464	60	—	2513	2513
43	515	—	515	61	—	2796	2796
44	572	—	572	62	—	3100	3100
45	635	—	635	63		3426	3426
46	704	—	704	64	—	3774	3774
47	779	—	779	65	—	4143	4143
48	860	—	860	65	—	4534	4534
49	948	—	948	67		4946	4946
50	1041	866	1037	68	—	5380	5380
51	1141	934	1121	69	—	5835	5835
52	1247	1023	1201	70	—	6312	6312

параболы и общую область определения двух кривых, а также приложив немного умения и терпения, можно в конце концов прийти к достаточно гладкой кривой.

Литература: [96, с. 239—247].

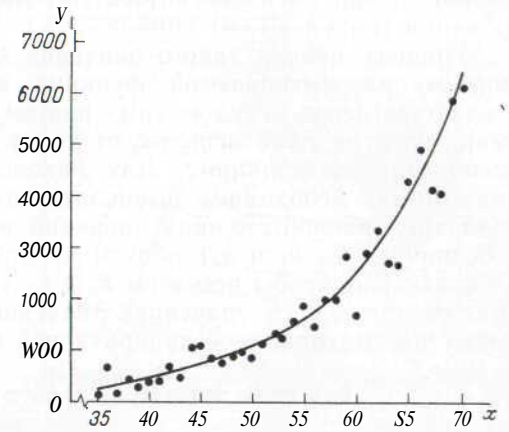


Рис. 7.4.1. Согласование двух парабол с помощью кривой синусов

7.5. ОПТИМИЗАЦИЯ ЗНАЧЕНИЯ ФУНКЦИИ НЕСКОЛЬКИХ ПЕРЕМЕННЫХ, МЕТОД НАИСКОРЕЙШЕГО СПУСКА

Часто приходится сталкиваться с проблемой минимизации (или максимизации) некоторой функции нескольких переменных. Теоретически для этого необходимо всего лишь приравнять нулю частные производные[3] и решить полученные уравнения. Однако реализация такого подхода может оказаться весьма трудоемким делом, поэтому для решения данной проблемы было разработано множество численных методов. Здесь описан один из подобных методов — метод наискорейшего спуска.

Предположим, нам необходимо найти точку минимума функции $f(x, y, z)$, зависящей от трех переменных x, y и z. Мы начинаем процесс минимизации с исходной точки (x_0, y_0, z_0) и движемся и направлении наискорейшего спуска. Оказывается, что это направление совпадает с вектором, исходящим из точки $(x_0, y_0,$ го), координаты которого пропорциональны соответствующим трем частным производным в этой точке. В результате мы передвигаемся в точку

$$(x_1, y_1, z_1) = \left(x_0 - k\frac{\partial f}{\partial x},\ y_0 - k\frac{\partial f}{\partial y},\ z_0 - k\frac{\partial f}{\partial z}\right). \qquad (7.5.1)$$

Число k выбирается таким образом, чтобы точка $(x\backslash, y_1, z\backslash)$ соответствовала наименьшему значению функции на всем луче, идущем в избранном нами направлении. Далее эта процедура повто-

[3] См. параграф 1.6.

ряется, что приводит к улучшенным оценкам точки минимума: (x_2, y_2, z_2), (x_3, y_3, z_3) и т. д.

Если выражения, задающие частные производные, являются слишком сложными, то вместо непосредственного дифференцирования можно воспользоваться методами численного дифференцирования (см. гл. 6).

Процесс поиска такого значения k, которое соответствует минимуму рассматриваемой функции на выбранном луче, может оказаться очень трудоемким, поэтому обычно вместо истинного минимума на луче используют точку *минимума* соответствующей *квадратичной функции*[4]. Для нахождения этого «квадратичного минимума» необходимо вычислить значения функции K_0, K_1 и K_2 для трех равноотстоящих значений величины k (которые можно обозначить k_0, k_1 и k_2) вблизи предполагаемой точки минимума. Желательно, чтобы при этом k_0 и k_2 лежали по разные стороны от минимизирующего значения. Значение k, соответствующее минимуму рассматриваемой квадратичной функции, задается формулой

$$k_{\min} = k_0 + \frac{1}{2}(k_1 - k_0)\left(1 - \frac{K_1 - K_0}{K_2 - 2K_1 + K_0}\right). \qquad (7.5.2)$$

Приведенное описание метода наискорейшего спуска является, по существу, готовым алгоритмом, который легко реализовать с помощью соответствующей программы. Однако иногда он сходится медленно. Это может произойти, например, в том случае, когда график функции имеет вид некоторой длинной узкой долины (оврага). Если исходная точка выбрана вблизи одного из концов этой долины, то процедура минимизации может дать траекторию, идущую зигзагом от одного склона в исходной части долины к другому и лишь медленно спускающуюся вниз по долине к истинной точке минимума. Процедуру оптимизации можно улучшить следующей модификацией.

1. Примените формулы метода наискорейшего спуска (7.5.1) и (7.5.2) для нахождения точки P_1 по исходной точке P_0.

2. Воспользуйтесь формулами метода наискорейшего спуска для нахождения точек с четными индексами P_2, P_4, P_6 и т. д. по непосредственно предшествующим точкам с нечетными индексами (P_1, Рз, P_5 и т. д.).

3. Найдите точку P_3 по формуле минимизации квадратичной функции, которая соответствует линии, соединяющей P_0 и P_2. Таким образом,

$$(x_3, y_3, z_3) = (x_2 + k(x_2 - x_0),\ y_2 + k(y_2 - y_0),\ z_2 + k(z_2 - z_0)). \qquad (7.5.3)$$

4. Для нахождения точек с нечетными индексами P_{2n+1} $(n > 1)$ примените формулу минимизации квадратичной функции, которая соответствует прямой линии, соединяющей точки P_{2n} и P_{2n-3}.

[4] Эту точку находят, принимая предположение о том, что переменная $f(x_1, y_1, z_1)$ является квадратичной функцией k.

Необходимая формула может быть получена из (7.5.3) заменой индексов: 3 на $2n + 1$, 2 на $2n$ и 0 на $2n - 3$.

Траектория, соединяющая точки с четными индексами, имеет, как правило, зигзагообразный вид и многократно пересекает долину, в то время как переход к очередной точке с нечетным индексом смещает по долине эти колебания (рис. 7.5.1).

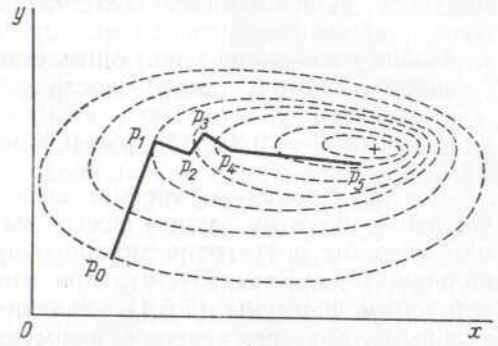


Рис. 7.5.1. Применение модифицированного метода наискорейшего спуска для нахождения точки минимума функции от двух переменных. (Приведены линии уровня функции, точка минимума указана крестом)

Пример 7.5.1. Найдем точку минимума функции

$$f(x, y) = \exp(x^2 + y^2) + \exp\{(x - 1)^2\} + \exp\{(y - 2)^2\}.$$

Частные производные этой функции таковы:

$$\partial f/\partial x = 2x \exp(x^2 + y^2) + 2(x - 1)\exp\{(x - 1)^2\},$$

$$\partial f/\partial y = 2y \exp(x^2 + y^2) + 2(y - 2)\exp\{(y - 2)^2\}.$$

Возьмем исходную точку

$$x_0 = 0{,}4;\quad \partial f/\partial x = 0{,}831\,951\,323\,4;\quad f = 7{,}341\,544\,519;$$

$$y_0 = 1{,}0;\quad \partial f/\partial y = 0{,}943\,302\,895\,3.$$

Если проверим значения $k = 0{,}1$ и $k = 0{,}05$ в формуле (7.5.1), то получим значения функции 7,417637066 и 7,322 035 734 соответственно. С помощью значений $k = 0$ (7,341 544 519), $k = 0{,}05$ (7,322 035 734) и $k = 0{,}1$ (7, 417 637 066) находим $k_{\min} = 0{,}033\,473\,966\,2$. Поэтому выбираем

$$x_1 = 0{,}372\,151\,289\,4;\quad df/dx = 0{,}321\,292\,99\,78;\quad f = 7{,}315\,486\,331;$$

$$y_1 = 0{,}968\,423\,910\,6;\quad df/dy = -0{,}297\,214\,523.$$

При проверке значений $k = 0{,}1$ и $k = 0{,}05$ получаем соответствующие значения функции 7,314395113 и 7,310418708. Используя значения функции при $k = 0$, $k = 0{,}05$ и $k = 0{,}1$, находим $k_{\min} = 0{,}053\,016\,406\,3$, откуда получаем

$$x_2 = 0{,}355\,117\,489\,3;\quad x_2 - x_0 = -0{,}044\,882\,510\,7;$$

$$y_2 = 0{,}984\,181\,156\,5;\quad y_2 - y_0 = -0{,}015\,8188435;\quad f = 7{,}310\,401\,834.$$

Теперь по формуле (7.5.3) найдем точку P_3. Проверяя значения $k = 0{,}5$ и $k = 0{,}25$, получаем соответствующие значения функции 7,310290573 и 7,309080967. С помощью этих значений и значения при $k = 0$ (7,310 401 834) получаем $k_{\min} = 0{,}255\,496\,057\,5$, откуда

$$x_3 = 0{,}343\,650\,184\,8; \quad \partial f/\partial x = 0{,}001\,826\,816\,1; \quad f = 7{,}309080\,443;$$

$$y_3 = 0{,}980\,139\,504\,4; \quad \partial f/\partial y = -0{,}006\,188\,022\,3.$$

После нескольких следующих шагов получаем решение, верное с точностью до пяти знаков после запятой:

$$x = 0{,}343\,45; \quad y = 0{,}980\,34; \quad f = 730908.$$

Данная процедура требует выполнения большого объема вычислений, поэтому лучше всего выполнять ее с помощью ЭВМ или хотя бы достаточно мощного программируемого настольного калькулятора. Заметьте, что при минимизации выражения в правой части формулы (7.5.1) значение $k_{\min}$ должно быть положительным, при применении формулы (7.5.3) $k_{\min}$ может иметь любой знак. Те же формулы могут быть полезными и для поиска максимума, в этом случае формула (7.5.2) дает точку максимума квадратичной функции $k_{\max}$, это значение должно быть отрицательным в формуле (7.5.1).

Литература: [19, с. 270—272; русский перевод с. 276—279], [89, с. 328—330].

7.6. ПРИЕМ, ПОЗВОЛЯЮЩИЙ УВЕЛИЧИТЬ ЗОНУ ХРАНЕНИЯ ДАННЫХ

Некоторые широко распространенные программируемые калькуляторы и мини-компьютеры имеют вполне достаточную зону памяти, выделенную для хранения программ, но весьма ограниченную зону хранения данных. Так, например, вычислительная машина Школы биологических наук при Университете Маккуори может выполнять программы длиной до 512 элементарных операций, в то же время зона данных состоит всего из 51 регистра. Однако в каждом из этих регистров может быть записано многозначное число (свыше двенадцати знаков в десятичной системе счисления).

В большинстве случаев исследователю в запоминаемых числах нужно лишь несколько первых значащих цифр (скажем, первые шесть), поэтому вполне оправдано хранение двух и даже трех чисел в одной ячейке. Этот метод хорошо известен пользователям ЭВМ первого поколения. Необходимы соответствующие обслуживающие программы, обеспечивающие преобразования чисел для отправки их на хранение и для обратной расшифровки. Поэтому данный метод неприменим на машинах с очень ограниченной зоной памяти, выделенной для хранения программ.

Пример 7.6.1. Рассмотрим некоторую вычислительную машину, в которой при записи чисел в память двенадцать разрядов выделено для мантиссы, один — для знака числа, два — для его характеристики и еще один — для знака характеристики. В такой машине число —0,000 123 хранится в памяти в виде —1,230 000 000 00 — 04.

Предположим, что нам необходимо хранить числа, лежащие в диапазоне от 300 до 900, и что для наших целей вполне достаточно запомнить числа с точностью до шести знаков в десятичной системе счисления. Так, например, пару чисел 379,6548 и 436,0917 можно хранить в виде одного числа с большим количеством знаков: 3,796 554 360 92 + 11. Этого можно достичь, используя простую подпрограмму, которая предполагает следующие действия:

возьмите первое число:	3,796 548 000 00 + 02
умножьте его на 10^3:	3,796 548 000 00 + 05
возьмите ближайшее целое число:	3,796 550 000 00 + 05
умножьте его на 10^6:	3,796 550 000 00 + 11 (A)
возьмите второе число:	4,360 917 000 00 + 02
умножьте его на 10^3:	4,360 917 000 00 + 05
возьмите ближайшее число:	4,360 920 000 00 + 05 (B)
сложите (A) и (B):	3,796 554 360 92 + 11

Когда необходимо вызвать данную пару чисел из памяти, два исходных числа выделяются с помощью следующей подпрограммы:

вызовите число из памяти:	3,796 554 360 92 + И (a)
разделите его на 10^6:	3,796 554 360 92 + 05
возьмите целую часть:	3,796 550 000 00 + 05 (b)
умножьте ее на 10^6:	3,796 550 000 00 + 11 (c)
вычтите (c) из (a):	4,360 920 000 00 + 05
разделите результат на 10^3:	4,360 920 000 00 + 02 (d)
разделите (b) на 10^3:	3,796 550 000 00 + 02 (e)

Первое число получено в строке (e), второе — в строке (d).

Описанная процедура, возможно, самая простая из процедур этого типа. Дальнейшее совершенствование ее может быть достигнуто использованием подходящего масштаба, заменой знака и применением более сложных правил хранения нескольких чисел в одной ячейке.

7.7. УПРАЖНЕНИЯ

1. а) По формуле (6.6.6) на основе данных табл. 7.2.1 найдите дневной прирост веса в контрольной группе крыс на 128-й день.

б) По формуле (7.2.1) на основе данных табл. 7.2.1 найдите дневной прирост веса в контрольной группе крыс на 128-й день.

в) Какой из этих двух результатов более надежен?

2. Найдите сглаженные значения в точках $x = 47$ и $x = 57$ по данным табл. 4.2.1 методом, описанным в примере 7.3.1. Через сглаженные точки, соответствующие значениям $x = 42$, 47, 52, 57 и 62, проведите сплайн-кривую.

3. С помощью кривой квадратов согласуйте две параболы, приведенные в примере 7.4.1.

4. Опишите процедуры записи в одну ячейку трех чисел и последующего восстановления их в первоначальном виде для машины из примера 7.6.1. Известно, что все рассматриваемые при этом числа лежат в диапазоне от —90 до 360.

Часть II. ОСНОВНЫЕ МЕТОДЫ СТАТИСТИКИ

8. ВЕРОЯТНОСТЬ, СТАТИСТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ И МОМЕНТЫ

В данной главе собраны основные результаты теории вероятностей и математической статистики, знание которых необходимо для чтения последующих глав. Вначале описаны аксиомы и операционные правила теории вероятностей; далее обсуждаются одномерные статистические распределения, их моменты, характеристики основной тенденции, дисперсия, асимметрия и эксцесс. В двух последних параграфах главы описаны двумерные распределения и моменты.

8.1. АКСИОМЫ И ОПЕРАЦИОННЫЕ ПРАВИЛА ТЕОРИИВЕРОЯТНОСТЕЙ

Игрок[1] бросает монету десять раз. Этот процесс можно рассматривать как некоторый *эксперимент.* При каждом подбрасывании выпадает либо «герб» (Г), либо «решетка» (Р). Конкретная последовательность результатов типа «герб» и «решетка» называется исходом эксперимента. Очевидно, что описанный сейчас эксперимент имеет $2^{10} = 1024$ возможных исхода. Например, одним из возможных исходов является последовательность ГГРГРРР **ГГГ.**

Игрок может подсчитывать число выпадений герба. Это число является *случайной величиной,* которая может принимать одно из следующих неотрицательных целых значений: О, 1, 2, . . ., 10.

Предположим, что получена последовательность, содержащая шесть выпадений герба и четыре выпадения решетки. В таком случае можно сказать, что произошло событие «шесть выпадений герба и четыре выпадения решетки», или событие «четное число выпадений герба», или событие «число выпадений герба превосходит число выпадений решетки».

Каждому событию A можно сопоставить некоторое число $P(A)$, неотрицательное и не превосходящее единицу. Это число называют *вероятностью* события. Сумма вероятностей всех взаимоисключающих событий, связанных с тем или иным экспериментом, равна единице. Так, в рассмотренном игровом примере вероятность четного числа выпадений герба и вероятность нечетного числа выпадений герба в сумме составляют единицу. Если событие A происходит в результате некоторого эксперимента с вероятностью $P(A)$ и этот эксперимент повторяют неограниченное число раз, то доля тех экспериментов, в результате которых произошло событие A, будет равна $P(A)$.

Теперь обратим внимание на следующие аксиомы и операционные правила теории вероятностей:

$$0 \leqslant P(A) \leqslant 1, \tag{8.1.1}$$

$$P(\text{не Л}) = 1 - P(A), \tag{8.1.2}$$

$$P(A_1 \text{ или } A_2 \text{ или } \ldots \text{ или } A_n) = P(A_1) + \ldots + P(A_n),$$

если $A_1, A_2, \ldots, A_n$—взаимно исключающие события, (8.1.3)

$$P(A_1 \text{ и } A_2 \text{ и } \ldots \text{ и Л}_n) = P(A_1)\,P(A_2) \ldots P(A_n),$$

если $A_1, A_2, \ldots, A_n$—взаимно независимы. **(8.1.4)**

Условную вероятность события B при условии, что произошло событие A, обозначают $P(B|A)$. Заметим, что

$$P(A \text{ и } B) = P(A)\,P(B|A). \tag{8.1.5}$$

Если A и B независимы, то

$$P(B|A) = P(B). \tag{8.1.6}$$

Из формулы (8.1.6) видно, что формула (8.1.5) сводится к (8.1.4) в том случае, когда A и B независимы.

Пример 8.1.1. Игрок бросает пять раз симметричную (хорошо сбалансированную) монету. Считая, что подбрасывания независимы, определите, какова вероятность того, что герб выпадает точно два раза?

Единственными исходами, при которых герб выпадает точно два раза, являются следующие десять последовательностей:

ГГРРР, ГРГРР, ГРРГР, ГРРРГ, РГГРР,
РГРГР, РГРРГ, РРГГР, РРГРГ, РРРГГ.

Вероятность выпадения герба при любом подбрасывании равна $\frac{1}{2}$, этому же числу равна и вероятность выпадения решетки. Пять подбрасываний монеты независимы. В соответствии с формулой (8.1.4) заключаем, что вероятность получить Г при первом подбрасывании, Г при втором, Р при третьем, Р при четвертом и Р при пятом равна: $\left(\frac{1}{2}\right)^5 = \frac{1}{32}$. Для каждой из остальных девяти последовательностей вероятность также равна $\frac{1}{32}$. Каждая из этих десяти последовательностей содержит по два выпадения

[1] Теория статистики находит применение почти в каждой области человеческой деятельности. Однако удобно рассматривать основные понятия на примере ситуаций, возникающих в азартных играх.

герба, и эти последовательности являются взаимно исключающими. Из формулы (8.1.3) следует, что искомая вероятность представляет собой сумму этих десяти равных между собой вероятностей и она равна $\frac{10}{32}$.

Пример 8.1.2. Игрок независимым образом пять раз бросает симметричную монету. В параграфе 10.1 [2] мы увидим, что

$$P(0 \text{ гербов}) = \frac{1}{32}, \quad P(3 \text{ герба}) = \frac{10}{32},$$

$$P(1 \text{ герб}) = \frac{5}{32}, \quad P(4 \text{ герба}) = \frac{5}{32},$$

$$P(2 \text{ герба}) = \frac{10}{32}, \quad P(5 \text{ гербов}) = \frac{1}{32}.$$

Заметим, что каждая из этих вероятностей удовлетворяет неравенству (8.1.1). Более того, эти шесть событий являются взаимно исключающими и образуют полную группу событий, и сумма их вероятностей равна единице. Какова вероятность того, что герб выпадет по крайней мере два раза?

События «герб выпал два раза», «три раза», «четыре раза» и «пять раз» являются взаимно исключающими, и, следовательно, по формуле (8.1.3) вероятность того, что герб выпадет по крайней мере дважды, равна:

$$\frac{10}{32} + \frac{10}{32} + \frac{5}{32} + \frac{1}{32} = \frac{26}{32}.$$

С другой стороны, мы замечаем, что вероятность выпадения герба по крайней мере два раза в точности равна вероятности того, что не будет нулевого или однократного выпадения герба. В силу формулы (8.1.3) вероятность нулевого или однократного выпадения герба равна:

$$\frac{1}{32} + \frac{5}{32} = \frac{6}{32}.$$

Следовательно, в соответствии с формулой (8.1.2) вероятность того, что герб выпадет по крайней мере два раза, равна:

$$1 - \frac{6}{32} = \frac{26}{32}.$$

Пример 8.1.3. Найдем вероятность того, что игрок из примера 8.1.2 получит четное число выпадений герба при условии, что герб выпал по крайней мере два раза.

Обозначим событие «герб выпал по крайней мере два раза» буквой *A*, а событие «герб выпал четное число раз» — буквой *B*.

[2] См. также пример 8.1.1.

Нам необходимо найти вероятность $P(B|A)$; по формуле (8.1.5) можно записать

$$P(B|A) = P(A \text{ и } B)/P(A).$$

Из примера 8.1.2 известно, что $P(A) = \frac{26}{32}$. Для того чтобы произошли оба события *A и B,* нужно, чтобы герб выпал два или четыре раза. События «герб выпал два раза» и «герб выпал четыре раза» являются взаимно исключающими, поэтому в соответствии с формулой (8.1.3) мы складываем вероятности этих событий, чтобы определить вероятность того, что произойдет одно из этих событий.

$$P(A \text{ и } B) = P(\text{герб выпал два раза или четыре раза}) =$$

$$= \frac{10}{32} + \frac{5}{32} = \frac{15}{32}.$$

Окончательно

$$P(B|A) = \frac{15}{32} \Big/ \frac{26}{32} = \frac{15}{26}.$$

Литература: [3, с. 41—48], [7, с. 1—11; русский перевод с. 11—23], [26, с. 4—29], [45, с. 5, 33—52], [46, с. 4—15], [70, с. 6—52], [93, с. 99—121], [101, с. 1—26; русский перевод с. 11—37], [102, с. 58—92].

8.2. ДИСКРЕТНЫЕ И НЕПРЕРЫВНЫЕ РАСПРЕДЕЛЕНИЯ

Во многих экспериментальных ситуациях события, интересующие исследователя, являются взаимно исключающими; такие события можно представить неотрицательными целыми числами 0, 1, 2, ... (например, число выпадений герба в эксперименте с подбрасыванием монеты). Обозначим вероятность события j символом p_j $(I = 0, 1, 2, \ldots)$. Мы можем определить случайную величину A", которая принимает значение j, когда происходит событие j. Тогда

$$P(X = r) = p_r, \tag{8.2.1}$$

$$P(X \leqslant r) = \sum_{j=0}^{r} p_j, \tag{8.2.2}$$

$$P(r < X \leqslant s) = \sum_{i=r+1}^{s} P_j, \tag{8.2.3}$$

$$\sum_{j=0}^{\infty} p_j = 1. \tag{8.2.4}$$

Вероятности $\{p_j\}$ определяют *распределение вероятностей* на множестве неотрицательных целых чисел. Распределение является *дискретным,* так как случайная величина может принимать лишь определенные дискретные значения (в данном случае — неотрицательные целые числа). Сумму в формуле (8.2.2) иногда обозначают символом $F(r)$ и называют *функцией распределения.*

По-видимому, дискретные распределения на множестве неотрицательных целых чисел — наиболее типичный случай дискретных распределений, но встречаются и многие другие типы; например, распределения на множестве положительных целых чисел или на множестве всех целых чисел (в том числе и отрицательных). Типичные дискретные распределения описаны в гл. 10.

Некоторые случайные величины, такие, как рост и вес, по существу, непрерывны, им соответствуют *непрерывные распределения.* Для такой случайной величины X определяют кумулятивную *функцию распределения* $F(x)$, такую, что

$$P(X \leqslant x) = F(x). \tag{8.2.5}$$

Очевидно, что $F(x)$ не может убывать, если величина x растет и $F(\infty) = 1$. В случае когда функция $F(x)$ дифференцируема,

$$P(x < X \leqslant x + dx) = f(x)\,dx, \tag{8.2.6}$$

где

$$f(x) = \frac{d}{dx} F(x) \tag{8.2.7}$$

и

$$F(x) = \int_{-\infty}^{x} f(x)\,dx. \tag{8.2.8}$$

Функция $f(x)$ называется *функцией плотности вероятностей,* соответствующей данному распределению, она может принимать лишь неотрицательные значения. Заметим также, что

$$P(a < X \leqslant b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a). \tag{8.2.9}$$

Нижним пределом интеграла в формуле (8.2.8) является $-\infty$. Некоторые случайные величины, например рост, могут принимать лишь положительные значения, для таких случайных величин значения обеих функций $f(x)$ и $F(x)$ при отрицательных x должны быть нулевыми.

Несколько из наиболее важных непрерывных распределений описано в гл. 9.

Пример 8.2.1. Монету бросают до тех пор, пока впервые не выпадет герб. Если X означает необходимое для этого число подбрасываний, то X будет дискретной случайной величиной, принимающей значения на множестве целых положительных чисел.

Пример 8.2.2. Биологу нужно сравнить концентрации бактерий двух типов A и B. Он подсчитывает число бактерий каждого типа в чашке Петри. Обе эти случайные величины распределены на множестве неотрицательных целых чисел. Далее биолог вычисляет отношение значений этих двух случайных величин. Получаемое частное будет случайной величиной, распределенной на множестве

неотрицательных рациональных чисел (т. е. таких точек числовой прямой, которые лежат на луче неотрицательных чисел и координаты которых могут быть выражены отношением двух целых чисел.

Литература: [3, с. 55]. [26, с. 38—40, 59—64], [45, с. 73—85], [46, с. 26—46, 52—89], [70, с. 53—102], [101, с. 30—48; русский перевод с. 42—65], [102, с. 98—121].

8.3. СРЕДНЕЕ И ДИСПЕРСИЯ

Среднее, или *ожидаемое,* значение* случайной величины X, распределенной на множестве неотрицательных целых чисел, определяют выражением

$$\mu = \sum_{j=0}^{\infty} j p_j, \tag{8.3.1}$$

а для непрерывного распределения в диапазоне $(-\infty, \infty)$ с функцией плотности вероятностей $f(x)$ — выражением

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx. \tag{8.3.2}$$

Среднее представляет собой характеристику (меру) положения значений случайной величины (см. [3, с. 21—22], [93, с. 45—68]).

Дисперсия распределения является мерой разброса распределения относительно среднего значения (см. [3, с. 23], [93, с. 69—88]). В случае дискретного распределения на множестве неотрицательных целых чисел дисперсия задается выражением

$$\sigma^2 = \sum_{j=0}^{\infty} (j - \mu)^2 p_j, \tag{8.3.3}$$

а для непрерывного распределения в диапазоне $(-\infty, \infty)$ — выражением

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx. \tag{8.3.4}$$

В случае когда все наиболее вероятные исходы сконцентрированы вблизи среднего значения, дисперсия σ^2 мала. Часто пользуются другим (математически эквивалентным) вариантом формулы дисперсии:

$$\sigma^2 = \sum_{j=0}^{\infty} j^2 p_j - \mu^2 \quad \text{(дискретный случай)}, \tag{8.3.5}$$

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 \quad \text{(непрерывный случай)}. \tag{8.3.6}$$

---

* Это значение называют также математическим ожиданием величины X.— *Примеч. пер.*

Квадратный корень из дисперсии называют *стандартным отклонением* и обозначают символом a.

Литература: [7, с. 34—42; русский перевод с. 50—58], [9, с. 100—102], [16, с. 24—32], [45, с. 16—20, 80—83], [70, с. 103—121], [81, с. 32—54], [93, с. 45—88], [105, с. 19—21, 31—35].

8.4. ВЫБОРОЧНОЕ СРЕДНЕЕ И ВЫБОРОЧНАЯ ДИСПЕРСИЯ

В большинстве экспериментальных ситуаций мы не знаем среднее и дисперсию рассматриваемого нами распределения. Необходимо оценить значения этих величин исходя из имеющихся экспериментальных данных; для этой цели обычно используют величины *выборочного среднего* и *выборочной дисперсии* (см. гл. 13). Для выборки $x_1, x_2, \ldots, x_n$ объема n выборочное среднее $\bar{x}$ и выборочную дисперсию s^2 вычисляют следующим образом:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \tag{8.4.1}$$

$$m_2 = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \bar{x}^2, \tag{8.4.2}$$

$$s^2 = \{n/(n-1)\}\, m_2. \tag{8.4.3}$$

Величина m_2 является выборочным моментом второго порядка (см. параграф 8.6); следует отметить также сходство формул (8.3.5) и (8.4.2). В среднем оценка m_2 несколько занижена по сравнению со значением σ^2; именно для корректировки этого смещения в формулу для вычисления s^2 введен поправочный коэффициент $n/(n-1)$. *Выборочное стандартное отклонение* s есть квадратный корень из выборочной дисперсии.

Выборочные данные дискретных распределений и сгруппированные выборочные данные непрерывных распределений могут быть представлены в форме некоторых частот. Пусть значение x наблюдалось с частотой f_x, тогда формулы для вычислений выборочного среднего x и выборочного момента m_2 имеют вид

$$\bar{x} = \frac{1}{n}\sum_{\text{все } x} x f_x, \tag{8.4.4}$$

$$m_2 = \frac{1}{n}\sum_{\text{все } x} x^2 f_x - \bar{x}^2, \tag{8.4.5}$$

где

$$n = \sum_{\text{все } x} f_x \tag{8.4.6}$$

есть объем выборки.

Литература: [3, с. 21—24], [70, с. 144—147], [102, с. 34—57].

8.5. МЕДИАНА И МОДА

Медианным значением называют такое значение случайной величины, которое делит распределение на две равновероятные половины; как и среднее, медиана является мерой положения. В случае непрерывного распределения в диапазоне $(-\infty, \infty)$ медианой будет такое значение m, что

$$\int_{-\infty}^{m} f(x)\,dx = \int_{m}^{\infty} f(x)\,dx = \frac{1}{2}. \tag{8.5.1}$$

При дискретном распределении на множестве неотрицательных целых чисел для определения медианы необходимо ввести некоторые дополнительные уточнения. Условимся считать медианой m такое целое число, что

$$\sum_{j=0}^{m-1} p_j < \frac{1}{2} \text{ и } \sum_{j=0}^{m} p_j > \frac{1}{2}. \tag{8.5.2}$$

В некоторых распределениях удается найти такое целое число M, что

$$\sum_{j=0}^{M} p_j = \frac{1}{2}. \tag{8.5.3}$$

Тогда берут $m = M + \frac{1}{2}$. Если распределение симметрично, то среднее и медиана совпадают.

Можно определить и *выборочную медиану,* но для этого снова необходимо принять некоторые условия. В выборке объема $2N+1$ медианой считают наблюдение с номером $N+1$ при условии, что номера идут в порядке убывания соответствующих выборочных значений. Если объем выборки задается четным числом $2N$, то в качестве медианы берут полусумму N-го и $(N+1)$-го значений (в упорядоченном списке значений). *Мода* соответствует такому значению случайной величины, которое является точкой максимума для функции плотности. Обычно в статистике имеют дело с унимодальными распределениями, т. е. такими, у которых функции плотности имеют только одну точку максимума.

Целесообразно заметить, что при унимодальном распределении среднее, медиана и мода расположены именно в таком порядке (либо в обратном); таким образом, медиана лежит на числовой оси между средним и модой (и притом ближе к среднему).

Литература: [3, с. 21—22], [9, с. 100—101], [53, с. 38—40; русский перевод с. 62—65], [81, с. 1—55], [93, с. 45—88], [105, с. 19—25].

8.6. МОМЕНТЫ БОЛЕЕ ВЫСОКОГО ПОРЯДКА, АСИММЕТРИЯ И ЭКСЦЕСС

Моментом r-го порядка относительно начала координат называют в случае дискретного распределения на множестве неотрицательных целых чисел значения

$$\mu_r = \sum_{j=0}^{\infty} j^r p_j, \tag{8.6.1}$$

а в случае непрерывного распределения в диапазоне $(-\infty, \infty)$ – значение

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x)\, Ax. \tag{8.6.2}$$

Иногда эти моменты называют еще *нецентральными моментами.* Следует заметить, что средние значения, определяемые выражениями (8.3.1) и (8.3.2), являются моментами первого порядка относительно начала координат; таким образом,

$$\mu \equiv \mu_1'. \tag{8.6.3}$$

Более того, значения дисперсий, вычисляемые по формулам (8.3.5) и (8.3.6), получают вычитанием квадрата среднего из нецентрального момента второго порядка.

Моменты относительно среднего определяют следующим образом:

$$\mu_r = \sum_{j=0}^{\infty} (j-\mu)^r p_j \quad \text{(дискретный случай)}, \tag{8.6.4}$$

$$\mu_r = \int_{-\infty}^{\infty} (x-\mu)^r f(x)\, dx \quad \text{(непрерывный случай)}. \tag{8.6.5}$$

Эти моменты называют также *центральными моментами.* Отметим, что дисперсии, определяемые формулами (8.3.3) и (8.3.4), являются центральными моментами второго порядка:

$$\sigma^2 = \mu_2. \tag{8.6.6}$$

На практике моменты порядка выше четырех используют редко, однако в исследовании иногда применяют центральные моменты третьего и четвертого порядков (например, при подборе кривой пирсоновского типа, см. гл. 11). Для симметричных распределений все центральные моменты нечетного порядка равны нулю; они положительны, если распределение асимметрично и имеет длинный «хвост» справа от средней, и отрицательны, если подобное распределение имеет длинный «хвост» слева. Поэтому функция моментов

$$\sqrt{\beta_1} = \mu_3/(\mu_2)^{\frac{3}{2}} \tag{8.6.7}$$

часто служит мерой *асимметрии.* Центральные моменты четных порядков всегда положительны, индикатор *эксцесса* (или остроты пика функции плотности) задается выражением

$$\beta_2 = \mu_4/\mu_2^2. \tag{8.6.8}$$

Для нормального распределения (см. параграф 9.1) $\sqrt{\beta_1} = 0$ и $\beta_2 = 3$. Распределения с более высокими значениями β_2 обычно имеют более острый пик, чем график функции плотности нормального распределения, а распределения с меньшими значениями β_2 — более сглаженный пик (по сравнению с нормальным).

Центральные и нецентральные моменты связаны следующими соотношениями:

$$\mu_2 = \mu_2' - (\mu_1')^2, \tag{8.6.9}$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3, \tag{8.6.10}$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4; \tag{8.6.11}$$

очевидно, что формулы (8.3.5) и (8.3.6) являются частными случаями этих соотношений.

Для выборки $x_1, x_2, \ldots, x_n$ объема n выборочный нецентральный момент r-го порядка вычисляют по формуле

$$m_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r. \tag{8.6.12}$$

Данные дискретных распределений и сгруппированные данные непрерывных распределений могут быть приведены в форме частот. Если частота появления значения x обозначена символом f_x, то указанные выборочные моменты вычисляют по формуле

$$m_r' = \frac{1}{n}\sum_{\text{все } x} x^r f_x, \tag{8.6.13}$$

где

$$n = \sum_{\text{все } x} f_x \tag{8.6.14}$$

есть объем выборки (см. пример 8.6.1).

Для вычисления центральных моментов выборки применяют формулы (8.6.9) и (8.6.11), заменяя в каждом случае величину μ величиной m. Следует отметить, что

$$\bar{x} \equiv m_1', \tag{8.6.15}$$

$$s^2 \equiv [n/(n-1)]\, m_2. \tag{8.6.16}$$

Иногда вводят поправки, связанные с учетом эффектов группировки данных (см. параграф 11.3). Выборочные характеристики асимметрии и эксцесса задаются формулами

$$\sqrt{b_1} = m_3/(m_2)^{\frac{3}{2}}, \tag{8.6.17}$$

$$b_2 = m_4/m_2^2. \tag{8.6.18}$$

Пример 8.6.1. В первых двух столбцах табл. 11.2.1 приведены сгруппированные данные о частотах в соответствующей выборке. Значение характеристики асимметрии $\sqrt{b_1}$ вычисляется следующим образом:

$$m_1' = (17 \times 34 + 22 \times 145 + 27 \times 156 + \ldots)/1000 = 37{,}875;$$

$$m_2' = (17^2 \times 34 + 22^2 \times 145 + 27^2 \times 156 + \ldots)/1000 = 1626{,}075;$$

$$m_3 = (17^3 \times 34 + 22^3 \times 145 + 27^3 \times 156 + \ldots)/1000 = 77986{,}575;$$

$$m_2 = 1626{,}075 - (37{,}875)^2 = 191\,559\,375;$$

$$m_3 = 77\,986{,}575 - 3 \times 1626{,}075 \times 37{,}875 + 2 \times (37{,}875)^3 = 1888{,}36172;$$

$$\sqrt{b_1} = 1888{,}36\,172/(191{,}559\,375)^{\frac{3}{2}} = 0{,}712\,246\,085.$$

Литература: [9, с. 105—110], [53, с. 85—86; русский перевод с. 83—85], [66, с. 24—25], [93, с. 89—98]. [101, с. 72—86; русский перевод с. 86—89].

8.7. ДВУМЕРНЫЕ РАСПРЕДЕЛЕНИЯ

Рассмотрим две дискретные случайные величины X и Y, принимающие целые неотрицательные значения (например, такой парой величин могут быть данные о возрасте мужа и жены для случайно выбранной семьи при условии, что возраст фиксируется целым числом лет — по состоянию на последний день рождения). Вероятность

$$P(X = i \text{ и } Y = j)$$

может быть обозначена символом p_{ij}, значения $\{p_{ij}\}$ $(i, j = 0, 1, 2, \ldots)$ определяют в совокупности некоторое двумерное распределение. Сумма значений pij по всем i и j равна, конечно, единице. Следует отметить соотношения

$$P(X = i \text{ и } Y = j) = p_{ij}, \tag{8.7.1}$$

$$P(X = i) = \sum_{j=0}^{\infty} p_{ij}, \tag{8.7.2}$$

$$P(X = i \mid Y = j) = p_{ij} / \left(\sum_{k=0}^{\infty} p_{kj} \right). \tag{8.7.3}$$

Математическое ожидание X при условии, что $Y = j$, может быть получено умножением выражения (8.7.3) на i и последующим суммированием полученных произведений по всем i. В случае когда величины X и Y *независимы,*

$$p_{ij} = P(X = i)\, P(Y = j). \tag{8.7.4}$$

Для непрерывных случайных величин X и Y, распределенных в диапазоне $(-\infty, \infty)$, можно определить *функцию плотности их совместного распределения* $f(x, y)$, такую, что

$$P(x < X \leqslant x + dx \text{ и } y < Y \leqslant y + dy) = f(x, y)\, dx\, dy. \tag{8.7.5}$$

Двойной интеграл от $f(x, y)$ по всей области изменения x и y равен единице; заметим, что

$$P(x < X \leqslant x + ax) = \int_{x}^{x+dx} \left(\int_{-\infty}^{\infty} f(x, y)\, dy \right) dx, \tag{8.7.6}$$

$$P(x < X \leqslant x + dx \mid y < Y \leqslant y + dy) = (f(x, y)\, dx) / \left(\int_{-\infty}^{\infty} f(x, y)\, dx \right). \tag{8.7.7}$$

Математическое ожидание X при условии, что значение Y лежит в диапазоне от y до $y + dy$, может быть вычислено умножением выражения из правой части (8.7.7) на x и последующим интегрированием по всей области изменения x. В случае когда величины *X и Y независимы,*

$$f(x, y)dxdy = P(x < X \leqslant x + dx)\, P(y < Y \leqslant y + dy). \tag{8.7.8}$$

Наиболее важное для приложений двумерное нормальное распределение описано в параграфе 9.6.

Пример 8.7.1. В табл. 8.7.1 приведены данные о числе супружеских пар некоторой гипотетической совокупности по соотносительным возрастным группам мужа и жены. Числа, заключенные в скобки, показывают долю пар данной группы в общей совокупности (например, 0,193 79 = 21 643/111 682). Пусть из данной совокупности случайным образом выбрана некоторая супружеская

Таблица **8.7.1. Численность возрастных групп супружеских пар (по возрасту мужа и жены) для некоторой гипотетической совокупности**

Возраст жены, лет	Возраст мужа, лет			Всего
	15—29	30-44	45+	
15-29	21 643 (0,193 79)	5 304 (0,047 49)	976 (0,008 74)	27 923 (0,250 02)
30—44	3 905 (0,034 97)	35219 (0,31535)	6742 (0,060 37)	45 866 (0,41069)
45+	391 (0,003 50)	4 596 (0,041 15)	32 906 (0,294 64)	37 893 (0,339 29)
Всего	25 939 (0,232 26)	45 119 (0,403 99)	40 624 (0,363 75)	111 682 (1,000 00)

пара. Обозначим возраст мужа в этой паре буквой Я, а возраст жены — буквой W. Тогда

а) $P(H$ лежит в группе $45+$, W лежит в группе $15—29) =$ $= 0{,}00874$;

б) $P(W$ лежит в группе $15—29) = 0{,}250\,02$;

в) P (Я лежит в группе $45+) = 0{,}363\,75$;

г) $P(W$ лежит в группе $15—29|H$ лежит в группе $45+) =$ $= 0{,}008\,74/0{,}363\,75 = 0{,}024\,03$;

д) $P(H$ лежит в группе $45+|W$ лежит в группе $15—29) =$ $= 0{,}008\,74/0{,}250\,02 = 0{,}034\,96$.

Литература: [7, с. 45—53; русский перевод с. 63—68], [26, с. 41, 72—75].

8.8. КОВАРИАЦИЯ И ДИСПЕРСИЯ

Моменты одномерного распределения были описаны в параграфах 8.3 и 8.6. Можно определить также двумерные и многомерные моменты; мы ограничимся рассмотрением особого двумерного момента — ковариации.

Пусть заданы две дискретные случайные величины A" и Y, распределенные на множестве целых неотрицательных чисел. Обозначим средние значения X и Y символами μ_x и μ_y соответственно. *Ковариация* величин X и Y определяется выражением

$$\operatorname{cov}(X,\ Y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} (i-\mu_x)(J-\mu_y)\,p_{ij}. \qquad (8.8.1)$$

Если наблюдается тенденция возрастания значений, принимаемых величиной X, при росте значений Y (и наоборот), то ковариация положительна. Это может встретиться в случаях, подобных тому, который представлен в табл. 8.7.1. Если наблюдается тенденция к росту значений X или уменьшения значений Y (и наоборот), то ковариация будет отрицательна. Ковариация двух независимых величин равна нулю, однако было бы *неверным* утверждение об обязательной независимости случайных величин с нулевой ковариацией [3]. Из нулевого значения ковариации можно заключить лишь об отсутствии какой-либо *линейной* связи между двумя случайными величинами.

На практике для вычисления ковариации чаще применяют другую (математически эквивалентную) формулу:

$$\operatorname{cov}(X,\ Y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} ijp_{ij} - \mu_x\mu_y. \qquad (8.8.2)$$

Аналогичными выражениями можно определить ковариацию и для двух непрерывных случайных величин:

$$\operatorname{cov}(X,\ Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y)\,f(x,\ y)\,dx\,dy \qquad (8.8.3)$$

[3] Подобное утверждение справедливо лишь для особого случая двумерного нормального распределения (см. параграф 9.6).

или

$$\operatorname{cov}(X,\ Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dx\,dy - \mu_x\mu_y. \qquad (8.8.4)$$

Следует отметить, что cov $(X,\ Y) =$ cov $(Y,\ X)$ и что ковариация случайной величины X с ней самой равна var (X).

В *матрице ковариации* представлены ковариации всевозможных пар случайных переменных из некоторого набора. Например, для набора случайных величин X, Y и Z матрица ковариации имеет вид

$$\begin{pmatrix} \operatorname{var}(X) & \operatorname{cov}(X,\ Y) & \operatorname{cov}(X,\ ZY \\ \operatorname{cov}(Y,\ X) & \operatorname{var}(Y) & \operatorname{cov}(Y,\ Z) \\ \operatorname{cov}(Z,\ X) & \operatorname{cov}(Z,\ Y) & \operatorname{var}(Z) \end{pmatrix}.$$

Диагональные элементы этой матрицы всегда положительны и матрица симметрична.

Коэффициент корреляции величин X и Y обычно обозначают буквой p, его значение равно частному от деления ковариации X и F на произведение их стандартных отклонений:

$$\mathrm{p} = \operatorname{cov}(X,\ Y)/(\sigma_X\sigma_Y). \qquad (8.8.5)$$

Значение этого коэффициента лежит на отрезке от -1 до 1, он характеризует степень линейности связи между величинами X и Y. Если X и Y связаны строго линейно (например, $Y = 10X + 3$), то абсолютная величина p равна единице. Если большим значениям X соответствуют большие значения Y, то корреляция положительна; если большим значениям X соответствуют малые значения Y, то корреляция отрицательна. Нулевая корреляция *не* означает независимость [4] величин X и Y, но из этого следует отсутствие какой-либо *линейной* зависимости между X и Y (см. [8, с. 271], [105, с. 237]).

Для выборки $(x_1, y_1), \ldots, (x_n, y_n)$ объема n можно вычислить *выборочный смешанный* момент m_{11}:

$$m_{11} = \frac{1}{n}\sum_{i=1}^{n} x_iy_i \sim xy. \qquad (8.8.6)$$

Следует отметить сходство данной формулы с выражением (8.8.2). В среднем оценка m_{11} лежит немного ниже истинного значения ковариации. Смещение устраняют, умножая тц на поправочный **коэффициент** $n/(n-1)$; таким образом, *выборочную дисперсию* определяют выражением

$$\text{выборочная дисперсия} = \{n/(n-1)\}m_{11}. \qquad (8.8.7)$$

Иногда данные о двумерном распределении приведены в форме частот (см. пример **8.8.2).** Пусть исход (x, y) наблюдался $f_{x,y}$ раз. Тогда выборочный смешанный момент *тц* вычисляют по формуле

$$m_{11} = \frac{1}{n}\sum_x\sum_y xyf_{x,y} - \bar{x}\bar{y}, \qquad (8.8.8)$$

[4] См. сноску 3.

где

$$n = \sum_x \sum_y f_{x,y} \tag{8.8.9}$$

показывает суммарную частоту всех наблюдений (т. е. объем выборки).

Для вычисления *выборочного коэффициента корреляции* r необходимо разделить выборочную ковариацию на произведение выборочных стандартных отклонений. Таким образом,

$$r = (\text{выборочная ковариация})/(s_X s_Y). \tag{8.8.10}$$

Значения этого коэффициента лежат между -1 и 1. Для вычислительных целей часто оказывается более удобной следующая формула, эквивалентная предыдущей:

$$r = \frac{m_{11}}{\sqrt{m_2(X)\, m_2(Y)}} \tag{8.8.11}$$

$m_2(X)$ означает второй центральный момент величины X (см. параграфы 8.4 и 8.6).

Пример 8.8.1. Бросают правильную шестигранную кость, величина X показывает численное значение исхода. Если значение исхода нечетно, то величина У принимает значение «нуль», в противном случае — значение «единица». Определим значения среднего и дисперсии величин X и Y, их ковариации и коэффициента корреляции.

Таблица **8.8.1. Двумерное распределение (см. пример 8.8.1)**

y	x						Всего
	1	2	3	4	5	6	
0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{2}$
1	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{2}$
Всего	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Данное двумерное распределение представлено в табл. 8.8.1.

Производим вычисления:

$$\mu_X = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \ldots + 6 \times \frac{1}{6} = 3{,}5;$$

$$\mu_Y = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = 0{,}5;$$

$$\sigma_X^2 = \left(1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \ldots\right) - (3{,}5)^2 = 2{,}916;$$

$$\sigma_Y^2 = \left(0^2 \times \frac{1}{2} + 1^2 \times \frac{1}{2}\right) - (0{,}5)^2 = 0{,}25.$$

Далее с помощью (8.8.2) получаем:

$$\begin{aligned}\operatorname{cov}(X, Y) = {} & (1 \times 0) \times \frac{1}{6} + (0 \times 1) \times 0 + (2 \times 0) \times 0 + \\ & + (2 \times 1) \times \frac{1}{6} + (3 \times 0) \times \frac{1}{6} + (3 \times 1) \times 0 + \\ & + (4 \times 0) \times 0 + (4 \times 1) \times \frac{1}{6} + (5 \times 0) \times \frac{1}{6} + \\ & + (5 \times 1) \times 0 + (6 \times 0) \times 0 + (6 \times 1) \times \frac{1}{6} - \\ & - 3{,}5 \times 0{,}5 = 0{,}25.\end{aligned}$$

Из формулы (8.8.5) следует, что

$$\rho = 0{,}25/(2{,}916 \times 0{,}25)^{\frac{1}{2}} = 0{,}292\,77.$$

Пример 8.8.2. В некотором случайном эксперименте возможны следующие двумерные исходы (X, Y): (1,1), (1,2), (1,3), (2,1), (2,2) и (2,3). Было проведено пятьдесят семь наблюдений. Первый из исходов появился 12 раз, второй — 6, третий — 5, четвертый — 15, пятый — 11 и шестой — 8 раз. Эти результаты представлены в табл. 8.8.2. Подсчитаем значение выборочного коэффициента корреляции.

Таблица **8.8.2. Результаты эксперимента с двумерными исходами. Числа в таблице показывают наблюдавшиеся частоты различных исходов (X, Y)**

x	y			Всего
	1	2	3	
1	12	6	5	23
2	15	11	8	34
Всего	27	17	13	57

Выборочные средние $\bar{x}$ и $\bar{y}$ получаем, применяя формулу (8.4.4) к распределениям, представленным в столбце и строке «Всего» (соответственно для x и y):

$$\bar{x} = (1 \times 23 + 2 \times 34)/57 = 1{,}5965;$$

$$\bar{y} = (1 \times 27 + 2 \times 17 + 3 \times 13)/57 = 1{,}7544.$$

Те же столбец и строка используются и для получения значений центральных моментов $m_2(X)$ и $m_2(Y)$ по формуле (8.4.5):

$$m_2(X) = (1^2 \times 23 + 2^2 \times 34)/57 - \bar{x}^2 = 0{,}240\,69;$$

$$m_2(Y) = (1^2 \times 27 + 2^2 \times 17 + 3^2 \times 13)/57 - \bar{y}^2 = 0{,}641\,43.$$

Для получения значения выборочного смешанного момента применяется формула (8.8.8):

$$m_{11} = «1 \times 1) \times 12 + (2 \times 1) \times 15 + (1 \times 2) \times 6 + (2 \times 2) \times$$
$$\times 11 + (1 \times 3) \times 5 + (2 \times 3) \times 8\}/57 - \bar{x}\bar{y} = 0{,}023\,700.$$

Значение выборочного коэффициента корреляции подсчитывают по формуле (8.8.11):

$$r = 0{,}023\,700/(0{,}240\,69 \times 0{,}641\,43)^{\frac{1}{2}} = 0{,}0603.$$

Литература: [7, с. 55—57; русский перевод с. 70—72], [8, с. 268—272], [45, с. 187—192], [81, с. 132—144].

8.9. УПРАЖНЕНИЯ

1. Некто бросает (правильную) шестигранную кость и сообщает, что выпало четное число. Какова вероятность того, что выпала «шестерка»?

2. Игрок бросает пять раз независимым образом симметричную монету. Какова вероятность того, что герб выпадет хотя бы дважды при условии, что герб уже выпал четное число раз? *Указание.* См. примеры 8.1.2 и 8.1.3.

3. Кость несимметрична, и нам сообщили, что $P(\text{один}) = \frac{1}{21}$, $P(\text{два}) = \frac{2}{21}$, $P(\text{три}) = \frac{3}{21}$, $P(\text{четыре}) = \frac{4}{21}$, $P(\text{пять}) = \frac{5}{21}$ и $P(\text{шесть}) = \frac{6}{21}$.

После бросания кости нам сообщают, что исход не меньше чем три. Какова вероятность того, что исход превосходит четыре?

4. Пусть каждой грани правильной шестигранной кости соответствует в выигрыше число очков j, равное числу j на грани. Рассмотрим случайную величину, представляющую выигрыш игрока, который бросает эту кость. Определите среднее, дисперсию, асимметрию и эксцесс данной случайной величины.

5. В табл. 8.8.1 представлено некоторое двумерное распределение. Найдите условное среднее значение величины X, соответствующее значению $Y=0$. Найдите также условное среднее значение величины Y, соответствующее значению $X=3$.

9. НОРМАЛЬНОЕ И ДРУГИЕ, СВЯЗАННЫЕ С НИМ РАСПРЕДЕЛЕНИЯ

Нормальное, t-, F-распределения и распределение χ^2 играют чрезвычайно важную роль в математической статистике, что является прямым следствием центральной предельной теоремы. Все остальные из перечисленных распределений определяются с помощью нормального. В этой главе мы кратко изложим основные свойства каждого из этих распределений. Будут также описаны логарифмически-нормальное и многомерное нормальное распределения.

Случайные ошибки, или флуктуации, в результатах научных экспериментов следуют приблизительно нормальному распределению. Есть много оснований ожидать, что это так. *Центральная предельная теорема* утверждает, что функция распределения случайной величины, являющейся суммой n независимых, одинаково распределенных случайных величин со средним μ и дисперсией σ^2, приближается к функции нормального распределения со средним $n\mu$ и дисперсией $n\sigma^2$, когда значения n становятся достаточно большими[1]. Часто конкретная переменная, которую мы измеряем, есть результат комбинирования значений многих величин, которые мы не измеряем или не можем измерить. Отклонение полученных наблюдений от средних или ожидаемых значений есть, таким образом, сумма ряда положительных и отрицательных, часто сравнимых между собой по величине отклонений.

Нормальное распределение является непрерывным распределением, полностью определяемым своим средним значением μ и дисперсией σ^2. Оно определено на всей вещественной прямой. Функция плотности вероятностей имеет следующий вид:

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu)^2/\sigma^2\right\}. \qquad (9.1.1)$$

Вероятность того, что случайная величина X, имеющая нормальное распределение со средним μ и дисперсией σ^2, заключена в пределах от a до b, можно вычислить, взяв интеграл от выражения (9.1.1) по интервалу (a, b). Интеграл не может быть выражен через элементарные функции, поэтому необходимо использовать таблицы. Сначала может показаться, что таблицы нужны для всех возможных значений x и комбинаций значений μ и σ^2. Однако это не так. Стандартизованная нормальная случайная величина $Y = (X-\mu)/\sigma$ имеет нулевое среднее и единичную дисперсию, ее функция плотности вероятностей определяется выражением

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right). \qquad (9.1.2)$$

Из соотношения

$$P(a < X \leqslant b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \qquad (9.1.3)$$

где

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)\,dy, \qquad (9.1.4)$$

[1] Данная формулировка центральной предельной теоремы наиболее простая; существуют варианты сложнее.

видно, что для вычислений, связанных с нормально распределенными величинами, необходима всего лишь одна основная таблица вероятностей. Таблицы стандартного нормального распределения приведены на с. 327—330, график плотности вероятностей стандартного нормального распределения приведен на рис. 9.2.2. Кривая $\varphi(x)$ симметрична относительно $x = 0$, значения $\Phi(x)$ для отрицательных x могут быть найдены вычитанием $\Phi(|x|)$ из единицы.

Пример 9.1.1. Известно, что в некоторой стране рост взрослых мужчин приблизительно описывается нормальным распределением со средним значением 175,6 см и стандартным отклонением 7,63 см. Выбирается случайным образом взрослый мужчина. Какова вероятность того, что его рост лежит в пределах от 175 до 185 см?

Вычисляем разность

$$\Phi((185 - 175{,}6)/7{,}63) - \Phi((175 - 175{,}6)/7{,}63),$$

которая равна: $\Phi(1{,}232) - \Phi(-0{,}079) = 0{,}891\,02 - 0{,}46852 = 0{,}42250$. Искомая вероятность приблизительно равна 0,423.

Литература: [3, с. 78—84], [16, с. 64—74], [26, с. 68—71], [48, с. 40—94]; [50, с. 100—108; русский перевод с. 130—141], [65, с. 25—30], [66, с. 10—19], [70, с. 123—125], [81, с. 68—75], [86, с. 124—147], [101, с. 156—157], [102, с. 144—159], [105, с. 70—84].

9.2. РАСПРЕДЕЛЕНИЕ χ^2

Рассмотрим множество взаимно независимых нормальных случайных величин $\{X_j\}$ $(j = 1, 2, \ldots, \nu)$, каждая из которых имеет нулевое среднее и единичную дисперсию. Возьмем еще одну случайную величину Y, определяемую следующим образом:

$$Y = \sum_{i=1}^{\nu} X_i^2.$$

Тогда Y подчиняется распределению χ^2 с ν степенями свободы.

Из этого определения видно, что случайная величина, подчиняющаяся распределению χ^2, должна быть положительна. Ясно также, что если Y_1 имеет распределение χ^2 с ν_1 степенями свободы, Y_2 имеет распределение χ^2 с ν_2 степенями свободы, а Y_1 и Y_2 независимы, то $Y = Y_1 + Y_2$ имеет распределение χ^2 с $\nu = \nu_1 + \nu_2$ степенями свободы.

Можно доказать, что функция плотности вероятности распределения χ^2 с ν степенями свободы равна:

$$f(x) = \frac{1}{2^{\frac{1}{2}\nu}\,\Gamma\left(\frac{1}{2}\nu\right)} x^{\left(\frac{1}{2}\nu\right)-1} e^{-\frac{1}{2}x} \quad (0 \leqslant x < \infty). \qquad (9.2.1)$$

Определение гамма-функции $\Gamma(n)$ дано в примечании к параграфу 11.2. Графики функции плотности вероятностей (для 2, 4 и

N степеней свободы) приведены на рис. 9.2.1; таблицы распределения χ^2 Даны на с. 331—332. В них приведены такие значения x, что

$$P(\chi_\nu^2 > x) = \alpha,$$

ν (число степеней свободы) указывается в заголовке строки, а (вероятность, или площадь под кривой $f(x)$ справа от x) указывается в заголовке столбца.

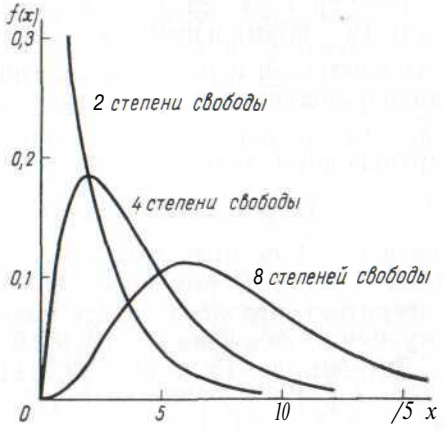


Рис. 9.2.1. Функции плотности вероятностей распределений χ^2 с 2, 4 и 8 степенями свободы. Кривая функции плотности распределения χ^2 пересекает ось y в точке (0; 0,5). Кривая функции плотности распределения χ^2 асимптотически приближается к оси y

Математическое ожидание распределения χ^2 с ν степенями свободы равно ν, а дисперсия равна 2ν. Распределение несиммет-

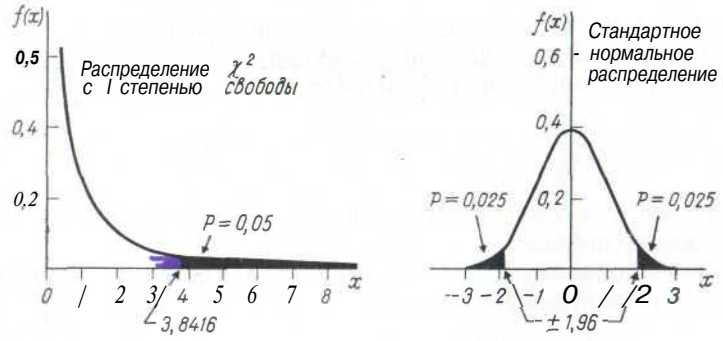


Рис. 9.2.2. Связь между распределениями χ_1^2 и стандартным нормальным распределением. Заштрихованная площадь под кривой равна сумме заштрихованных площадей под нормальной кривой; $3{,}8416 = (1{,}96)^2$

Обратим внимание на то, что если Y имеет распределение χ^2 с ν степенями свободы и ν достаточно велико (скажем, больше 30), то распределение величины $Z = \sqrt{2Y} - \sqrt{2\nu - 1}$ является приближенно нормальным с нулевым средним и единичной дис-

Персией. В этом случае можно применять таблицы нормального распределения.

Распределение χ^2 играет важную роль при исследовании выборочной дисперсии s^2 для выборки, взятой из нормально распределенной совокупности. Оно используется также в надежном приближенном критерии, предназначенном для проверки зависимости в таблицах сопряженности *.

Пример 9.2.1. Этот пример демонстрирует связь между стандартным нормальным распределением и распределением χ^2_1. В соответствии с ранее приведенным определением случайная величина, подчиняющаяся распределению χ^2_1, есть просто квадрат стандартной нормальной случайной величины. Обозначим стандартную нормальную случайную величину через u. Тогда

$$P(\chi^2 > a^2) = P(u < -a) + P(u > a) = 2P(u > a).$$

Когда a = 1,96, правая часть тождества равна 0,05. (Значение a^2 равно 3,8416.) С помощью таблиц распределения χ^2 можно удостовериться, что левая часть тождества также равна 0,05. Такая ситуация изображена на рис. 9.2.2.

Литература: [3, с. 90—92, 111], [7, с. 82; русский перевод с. 73—75], [26, с. 190—193], [45, с. 234—245], [46, с. 252—256], [48, с. 166—199], [50, с. 118; русский перевод с. 152—155], [53, с. 369—374; русский перевод с. 510—517], [70, с. 226—227], [75, с. 14].

9.3. t-РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА

Пусть X_1 имеет стандартное нормальное распределение, а X_2 — независимая от X_1 случайная величина, подчиняющаяся распределению χ^2 с v степенями свободы. Тогда

$$t = \frac{\sqrt{\nu} X_1}{\sqrt{X_2}}$$

определяется как случайная величина, имеющая t-распределение с v степенями свободы.

Можно доказать, что t_ν-распределение имеет следующую функцию плотности вероятностей:

$$f(x) = \frac{\Gamma\left(\frac{1}{2}\nu + \frac{1}{2}\right)}{\nu^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\nu\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \quad (-\infty < x < \infty). \tag{9.3.1}$$

Гамма-функция $\Gamma(n)$ определяется в примечании к параграфу 11.2. Графики функции плотности вероятности t_ν-распределения для некоторых значений v приведены на рис. 9.3.1; таблицы t-распределения даны на с. 333.

t-распределение симметрично относительно среднего, равного нулю; его дисперсия равна $\nu/(\nu - 2)$. t-распределение приобретает важность в тех случаях, когда рассматриваются выборочные средние и нет дисперсии генеральной совокупности. В этой ситуации мы вынуждены использовать выборочную дисперсию и соответственно t-распределение.

Для больших значений v дисперсия случайной величины t приблизительно равна единице; среднее равно нулю. Фактически для больших значений v (скажем, $\nu > 30$) распределение стремится к нормальному с нулевым средним и единичной дисперсией

Рис. 9.3.1. Функции плотности вероятностей t-распределения с 1, 5 и бесконечным числом степеней свободы. Последняя совпадает с плотностью стандартного нормального распределения

(в этом случае велика вероятность того, что значение X_2/ν близко к единице и, следовательно, t почти идентично X_1).

Пример 9.3.1. Найдем вероятность того, что случайная величина t с десятью степенями свободы лежит в пределах от 1,5 до 2,0. В соответствии с таблицами t-распределения

$$P(1{,}372 < t_{10} < 1{,}812) = 0{,}05,$$
$$P(1{,}812 < t_{10} < 2{,}228) = 0{,}025.$$

Требуемая вероятность приблизительно равна:

$$0{,}05 \times (1{,}812 - 1{,}5)/(1{,}812 - 1{,}372) + 0{,}025 \times$$
$$\times (2{,}0 - 1{,}812)/(2{,}228 - 1{,}812) = 0{,}047.$$

Литература: [3, с. 121—144], [16, с. 75—77], [45, с. 146—148], [46, с. 257—260], [49, с. 94—124], [50, с. 119; русский перевод с. 115—118], [53, с. 374—377; русский перевод с. 517—521], [70, с. 233].

* См. параграф 12.9.— *Примеч. пер.*

9.4. *F*-РАСПРЕДЕЛЕНИЕ

Пусть X_1 и X_2 — независимые случайные величины, имеющие распределение χ^2 с m и n степенями свободы соответственно. Тогда определим положительную случайную величину F с m и n степенями свободы:

$$F_{m,n} = \frac{(X_1/m)}{(X_2/n)}.$$

Можно доказать, что $F_{m,n}$-распределение имеет следующую функцию плотности вероятностей:

$$f(x) = \frac{\Gamma\left(\frac{1}{2}m + \frac{1}{2}n\right)}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \left(\frac{m}{n}\right)^{\frac{1}{2}m} \frac{x^{\left(\frac{1}{2}m\right)-1}}{\left(1+\frac{m}{n}x\right)^{\frac{1}{2}(m+n)}}$$

$$(0 \leqslant x < \infty). \qquad (9.4.1)$$

Гамма-функция $\Gamma(n)$ определена в примечаниях к параграфу 11.2. График функции плотности $F_{10,12}$-распределения показан на рис. 9.4.1, а таблицы $F_{m,n}$-распределения приведены на с. 334.

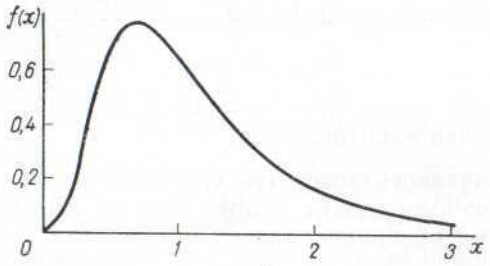


Рис. 9.4.1. Функция плотности вероятностей $F_{10,12}$-распределения

Математическое ожидание и дисперсия $F_{m,n}$-распределения равны $n/(n-2)$ и $\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ $(n>4)$ соответственно. Распределение несимметрично.

В пределе при $n \to \infty$ математическое ожидание $F_{m,n}$ становится равным единице, а дисперсия равна $2/m$. Эти величины представляют собой моменты случайной величины χ^2_m/m; фактически $mF_{m,\infty}$ имеет распределение χ^2 с m степенями свободы.

F-распределение приобретает особую важность, когда сравниваются выборочные дисперсии из нормально распределенных совокупностей. Оно также широко используется в регрессионном и дисперсионном анализе.

Пример 9.4.1. Этот пример иллюстрирует связь между распределением χ^2_m и $F_{m,\infty}$-распределением. Было показано, что

$$P\left(F_{m,\infty} > \frac{a}{m}\right) = P(\chi^2_m > a).$$

Пусть $m = 10$ и $a = 18{,}31$. Правая часть равна 0,05, легко убедиться, что и левая часть также равна 0,05.

Пример 9.4.2. Этот пример показывает связь между t_n- и χ^2_n-распределениями. В соответствии с определением (см. параграф 9.3) t_n есть отношение единичной нормальной случайной величины к квадратному корню из χ^2_n/n (все случайные величины независимы). Если мы возведем в квадрат t_n, то получим отношение $\chi^2_1/1$ к χ^2_n/n. Это отношение имеет $F_{m,n}$-распределение. Таким образом,

$$P(F_{1,n} > a^2) = P(t_n > a) + P(t_n < -a).$$

Положим $n = 10$ и $a = 2{,}23$, так что $a^2 = 4{,}97$. Правая часть равна 0,05, и мы можем убедиться, что и левая часть также равна 0,05.

Пример 9.4.3. Таблицы в приложении дают только правые хвостовые точки для F-распределения; распределение несиммет-[illegible]к). В этом примере покажем, как могут быть получены из таблиц левые хвостовые точки распределения. Найдем такое x, что $P(F_{7,9} < x) = 0{,}05$.

Из определения $F_{m,n}$ видно, что

$$P(F_{7,9} < x) = P\left(F_{9,7} > \frac{1}{x}\right).$$

В соответствии с таблицами $P(F_{9,7} > 3{,}68) = 0{,}05$ и требующееся значение x равно: $x = 1/3{,}68 = 0{,}272$. Было применено следующее правило для нахождения нижней 100α %-ной границы распределения: она равна величине, обратной верхней 100α %-ной границе $F_{n,m}$-распределения с тем же значением a.

Литература: [3, с. 145—150], [26, с. 202—213], [45, с. 269—870], [49, с. 75—89], [50, с. 123: русский перевод с. 162], [53, с. 377—382; русский перевод с. 521—526], [70, с. 231—232], [75, с. 14].

!).5. ЛОГАРИФМИЧЕСКИ-НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Пусть X — случайная величина. Говорят, что X имеет логарифмически-нормальное распределение, если существует такое число θ, что величина $Z = \ln(X - \theta)$ распределена нормально. Так как логарифмическая функция определена только для положительных значений X должны быть больше 6. Распределение случайной величины X включает три параметра: θ, μ и σ^2, последние два

являются средним и дисперсией нормального распределения. Плотность вероятности величины X имеет вид

$$f(x) = \frac{1}{(x-\theta)\,\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\{\ln(x-\theta)-\mu\}^2/\sigma^2\right]. \qquad (9.5.1)$$

Отметим также, что среднее и дисперсия X равны:

$$\text{среднее} = 0 + \exp\left(V + \frac{1}{2}\sigma^2\right), \qquad (9.5.2)$$

$$\text{дисперсия} = e^{2\mu}e^{\sigma^2}(e^{\sigma^2}-1). \qquad (9.5.3)$$

Распределение имеет одну моду и положительную скошенность. Когда σ^2 мало, то распределение по форме близко к нормальному.

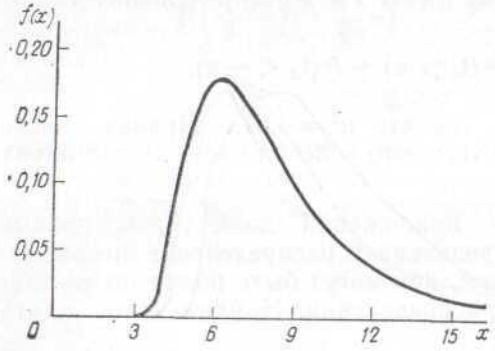


Рис. 9.5.1. Функция плотности вероятностей логарифмически-нормального распределения с параметрами $\theta = 3$, $\mu = 1{,}5$ и $\sigma^2 = 0{,}36$. Среднее равно 8,366, дисперсия равна 12,475

График функции плотности вероятностей логарифмически-нормального распределения с параметрами $0 = 3$, $\mu = 1{,}5$ и $\sigma^2 = 0{,}36$ приведен на рис. 9.5.1.

Известно, что во многих приложениях в равно нулю, и тогда X имеет двупараметрическое логарифмически-нормальное распределение. Это двупараметрическое распределение обладает по крайней мере одним преимуществом. Оно дает более реалистическое описание совокупностей таких положительно-определенных величин, как рост, вес и т. п., чем нормальное распределение, приписывающее положительные вероятности и отрицательным значениям переменной.

Методы оценивания параметров распределений рассмотрены в параграфе 13.25.

Пример 9.5.1. Найдем верхнюю 2,5 %-ную точку логарифмически-нормального распределения с параметрами 8, μ и σ^2.

Мы должны найти такое x, что $P(X < x) = 0{,}025$, или

$$P\left(\frac{\ln(X-\theta)-\mu}{\sigma} < \frac{\ln(x-6)-\mu}{\sigma}\right) = 0{,}025.$$

Случайная величина $\{\ln(X-6)-\mu\}/\sigma$ имеет стандартное нормальное распределение с верхней 2,5 %-ной точкой, равной

$(-1{,}96)$. Приравнивая $\{\ln(x-\theta)-\mu\}/\sigma$ к $(-1{,}96)$, мы получаем

$$x = \theta + \exp(\mu - 1{,}96\sigma).$$

Литература: [9, с. 153], [16, с. 78—79], [48, с. 112—136], [103, С. 5—16].

9.6. ДВУМЕРНОЕ И МНОГОМЕРНОЕ НОРМАЛЬНЫЕ РАСПРЕДЕЛЕНИЯ

Говорят, что две случайные величины X_1 и X_2 распределены по двумерному нормальному закону, если все линейные комбинации вида $a_1X_1 + a_2X_2$ являются нормально распределенными случайными величинами. Функция плотности совместного распределения величин X_1 и X_2 имеет вид

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} \exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right]. \qquad (9.6.1)$$

Параметры μ_1 и μ_2 равны математическим ожиданиям величин X_1 и X_2, σ_1^2 и σ_2^2 — дисперсии X_1 и X_2, а ρ — коэффициент корреляции. Ковариация величин X_1 и X_2 есть $\rho\sigma_1\sigma_2$. Когда коэффициент корреляции равен нулю, двумерная плотность, определенная в (9.6.1), равна произведению двух одномерных нормальных плотностей. Отсюда следует, что две нормально распределенные случайные величины независимы, если они не коррелируют *.

Условное математическое ожидание случайной величины X_2 при фиксированном X_1 есть линейная функция от X_1. Обозначив эту величину через $E(X_2|X_1)$, можно записать:

$$E(X_2|X_1) = \mu_2 + \rho(\sigma_2/\sigma_1)(X_1 - \mu_1). \qquad (9.6.2)$$

Ковариационная матрица величин X_1 и X_2 состоит из ковариации X_1 с X_1, ковариации X_1 с X_2, ковариации X_2 с X_1 и ковариации X_2 с X_2. Обозначим эту матрицу через V. Тогда

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Определитель матрицы V равен:

$$|\mathbf{V}| = \sigma_1^2\sigma_2^2(1-\rho^2).$$

Обратная матрица $\mathbf{V}^{-1}$ может быть легко вычислена с помощью алгебраических дополнений (см. параграф 1.10).

* Заметим, что только в случае нормально распределенных величин условие $\rho = 0$ необходимо и достаточно для независимости этих величин.— *Примеч. пер.*

Если мы под *{Xi}*, *{Xi}* и $\{\mu_i\}$ будем понимать вектор-столбцы X, x и $\boldsymbol{\mu}$ соответственно, то функция плотности вероятностей величины X, приведенная в (9.6.1), может быть записана следующим образом:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{V}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \mathbf{V}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}, \quad (9.6.3)$$

где $n = 2$. Формула (9.6.3) дает выражение для плотности n-мерного нормального распределения. Ковариационная матрица V имеет размерность $n \times n$, а вектор-столбцы X, x и $\boldsymbol{\mu}$ — размерность n.

Литература: [7, с. 397—416; русский перевод с. 366—384], [26, с. 75—77], [50, с. 129—132; русский перевод с. 167—170], [70, с. 198—218], [101, с. 158—169; русский перевод с. 173—184].

9.7. УПРАЖНЕНИЯ

1. Известно, что в некоторой стране рост взрослых мужчин имеет приближенно нормальное распределение со средним значением 175,6 см и стандартным отклонением 7,63 см. Случайным образом выбрано трое взрослых мужчин. Какова вероятность того, что все трое выше 185 см?

2. Известно, что в некоторой стране рост взрослых мужчин распределен приближенно нормально со средним значением 175,6 см. Дисперсия неизвестна. Мы знаем, однако, что 10 % взрослых мужчин имеют рост выше 187 см. Вычислите дисперсию.

3. Вычислите вероятность того, что случайная величина χ^2_{10} больше 20.

4. Вычислите вероятность того, что случайная величина χ^2_{60} больше 100.

5. Найдите вероятность того, что случайная величина t_{10} лежит в пределах от (—1) до 2.

6. Найдите верхнюю и нижнюю 2,5 %-ные точки для распределения $F_{19,59}$.

10. ОСНОВНЫЕ ДИСКРЕТНЫЕ РАСПРЕДЕЛЕНИЯ

В этой главе приводится характеристика наиболее распространенных дискретных распределений (биномиальное, пуассоновское, геометрическое, отрицательное биномиальное, логарифмическое и гипергеометрическое). Мы также опишем ряд соотношений между распределениями и некоторые полезные аппроксимации распределений.

10.1. БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Когда некоторая монета подбрасывается вверх, результатом является выпадение герба с вероятностью p и выпадение решетки с вероятностью $q = 1 - p$. Вероятность получить r раз герб и $(n - r)$ раз решетку при n независимых подбрасываниях монеты равна:

$$P(\text{герб выпадает } r \text{ раз при } n \text{ подбрасываниях монеты}) = \binom{n}{r} p^r q^{n-r}. \quad (10.1.1)$$

Эта вероятность определяет *биномиальное распределение*. Частный случай (10.1.1) был доказан в примере 8.1.1, доказательство в общем случае аналогично. Комбинаторная система обозначений излагается в параграфе 1.5. Биномиальное распределение имеет

$$\text{среднее} = np, \quad (10.1.2)$$

$$\text{дисперсию} = npq. \quad (10.1.3)$$

На рис. 10.1.1 приведен график плотности биномиального распределения с параметрами $n = 10$, $p = 0{,}3$.

Вычисление отдельных биномиальных вероятностей и совокупной суммы вероятностей — весьма трудоемкая процедура. Обширные таблицы, предназначенные для этих целей, подготовлены Национальным бюро стандартов [71], Управлением материального обеспечения армии США [98] и Вайнтраубом [99]. Читателю ре-

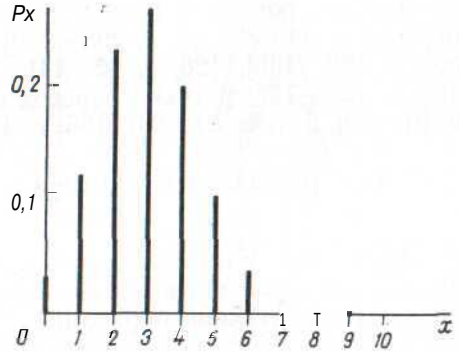


Рис. 10.1.1. Биномиальное распределение с параметрами $n = 10$ и $p = 0{,}3$

комендуется обратить особое внимание на нормальную аппроксимацию и F-метод, описанные соответственно в параграфах 10.2 и 10.3.

Методы оценки параметров распределения рассмотрены в параграфе 13.5.

Пример 10.1.1. Шесть коров могут бродить случайным образом по большому лугу, равномерно покрытому травой. Коровы не имеют склонности к скоплению, и каждая корова равно предпочитает любую точку луга. Пусть площадь луга равна Л, и коровы двигаются совершенно независимо одна от другой; найти вероятность того, что по крайней мере четыре коровы окажутся внутри квадрата с площадью a, расположенного в пределах луга.

Вероятность того, что какая-либо конкретная корова окажется внутри указанного квадрата, равна $p = a/A$. Тогда вероятность того, что четыре или более коров окажутся внутри данного квадрата, равна:

$$\binom{6}{4} p^4 q^2 + \binom{6}{5} p^5 q^1 + \binom{6}{6} p^6 q^0.$$

Пример 10.1.2. Известно, что в некоторой стране рост взрослых мужчин имеет приблизительно нормальное распределение со средним 175,6 см и стандартным отклонением, равным 7,63 см. Предположим, что мы выберем случайным образом 6 человек среди взрослых мужчин. Какова вероятность того, что не меньше четырех из них будут иметь рост между 175 и 185 см?

Из примера 9.1.1 мы знаем, что с вероятностью 0,423 рост случайно выбранного взрослого мужчины лежит между 175 и 185 см. Вероятность того, что не меньше четырех человек из шести будут иметь рост между 175 и 185 см, равна:

$$\binom{6}{4}(0{,}423)^4(0{,}577)^2+\binom{6}{5}(0{,}423)^5(0{,}577)^1+$$

$$+\cdot\left(\quad\right)(0{,}423)^6(0{,}577)^0=0{,}212.$$

Литература: [2, с. 6—20], [3, с. 63—69], [7, с. 30—31; русский перевод с. 34—36], [8, с. 295—299], [21, с. 135—141; русский перевод с. 152—158], [26, с. 42—44], [45, с. 91—99], [47, с. 50—86], [53, с. 120—125; русский перевод с. 171—178], [66, с. 26—31], [70, с. 64—69], [81, с. 84—94], [86, с. 112—118], [102, с. 122—132].

10.2. НОРМАЛЬНАЯ АППРОКСИМАЦИЯ БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Пусть монету подбрасывают n раз. Вероятность того, что число выпадений герба расположено между r_1 и r_2 (включительно), вычисляется следующим образом:

$$\sum_{r=r_1}^{r_2}\binom{n}{r}p^r q^{n-r}. \qquad (10.2.1)$$

Для больших n вычислять эту сумму довольно утомительно. Однако когда n больше 20, центральная предельная теорема [1] позволяет заменить эту сумму выражением

$$\int_{\alpha_1}^{\alpha_2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)dx, \qquad (10.2.2)$$

где

$$\alpha_1=\left(r_1-np-\frac{1}{2}\right)/(npq)^{\frac{1}{2}}, \qquad (10.2.3)$$

$$\alpha_2=\left(r_2-np+\frac{1}{2}\right)/(npq)^{\frac{1}{2}}. \qquad (10.2.4)$$

[1] См. параграф 9.1. Исходом биномиального эксперимента, включающего n испытаний, является сумма исходов каждого из n независимых испытаний. Приближенная формула для вычисления вероятности (10.2.1) была одной из самых ранних форм центральной предельной теоремы.

Площадь под кривой, определяемую (10.2.2), можно получить по таблицам нормального распределения.

Поправками $\pm\frac{1}{2}$ в выражениях для α_1 и α_2 часто пренебрегают, особенно когда n очень велико. Смысл введения поправок очевиден из рис. 10.2.1: нам требуется сумма биномиальных ординат; эта сумма равна площади многоугольника, которая аппроксимируется площадью, ограниченной нормальной кривой и ординатами точек α_1 и α_2.

Будем использовать случайную переменную X, чтобы обозначать число выпадений герба при n подбрасываниях. Тогда, используя (10.2.2), можем сказать, что $(X-np)/\sqrt{npq}$ приближенно

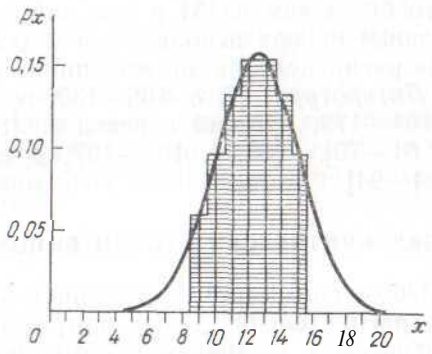


Рис. 10.2.1. Нормальное приближение биномиального распределения, $n=25$, $p=0{,}5$. Вероятность того, что значения случайной величины заключены между 9 и 15 включительно, равна сумме соответствующих ординат. Эта сумма равна заштрихованной площади гистограммы, значение которой приближенно равно площади под нормальной кривой между ординатами 8,5 и 15,5

подчиняется нормальному распределению с нулевым средним и единичной дисперсией. Из параграфа 9.2 следует, что $(X-np)^2/(npq)$ приближенно подчиняется распределению χ^2 с одной степенью свободы. Заметим, что $(X-np)^2$, представляющее собой квадрат разности между наблюдаемым и ожидаемым числом выпадений герба, также равно квадрату разности между наблюдаемым и ожидаемым числом выпадений решетки. Кроме того,

$$1/(npq)\equiv 1/(np)+1/(nq).$$

Из этого следует, что сумма

$$\frac{(\text{число наблюдаемых гербов}-\text{число ожидаемых гербов})^2}{\text{число ожидаемых гербов}}+$$

$$+\frac{(\text{число наблюдаемых решеток}-\text{число ожидаемых решеток})^2}{\text{число ожидаемых решеток}} \qquad (10.2.5)$$

подчиняется распределению χ^2 с одной степенью свободы.

Пример 10.2.1. Симметричную монету подбрасывают 1000 раз. Какова вероятность того, что число выпадений герба будет отклоняться от ожидаемого числа, равного 500, более чем на 31?

Используя симметричность нормальной кривой и аппроксимацию биномиального распределения нормальным (без поправок), найдем требуемую вероятность:

$$2\left\{1-\Phi\left(\frac{531-500}{\sqrt{250}}\right)\right\}=0{,}05.$$

С другой стороны, можно вычислить

$$\chi_1^2=\frac{312}{500}+\frac{31^2}{500}=3{,}84$$

и в соответствии с таблицей найти, что вероятность получить такое или большее значение χ_1^2 равна 0,05.

Заметим, что по определению метод хи-квадрат позволяет найти вероятность отклонения числа выпадений герба от ожидаемого более чем на 31 в том или другом направлении. Когда нормальная кривая используется в задачах такого типа, то оба «хвоста» распределения должны приниматься во внимание.

Литература: [7, с. 136—139; русский перевод с. 126—130], [21, с. 164—179; русский перевод с. 184—199], [45, с. 105—111], [47, с. 61—70], [53, с. 106—107; русский перевод с. 153—154], [81, с. 84—94].

10.3. КРИТИЧЕСКИЕ ТОЧКИ БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Обозначим через X случайную величину, имеющую биномиальное распределение с параметрами n и p. Верхним ЮОа %-ным критическим значением биномиального распределения с параметрами n и p называется целое число r, такое, что

$$P(X\geqslant r)\leqslant\alpha<P(X\geqslant r-1). \qquad (10.3.1)$$

Для определения критических значений можно использовать таблицы биномиального распределения. Приближенно критические точки можно вычислить с помощью нормальной аппроксимации, описанной в параграфе 10.2. Однако есть и другие способы вычисления критических значений. Обратим внимание на то, что

$$P(X\geqslant r)=P\left(Y\geqslant\frac{r}{n+1-r}\,\frac{1-p}{p}\right), \qquad (10.3.2)$$

где случайная величина Y имеет F-распределение с $2(n+1-r)$ и $2r$ степенями свободы (см. [7, с. 148]). Поэтому для нахождения верхней 100α %-ной точки биномиального распределения можно применить следующую процедуру:

1. Вычислить r_0 — нормальное приближение ЮОа %-ной точки и округлить значение r_0 до ближайшего целого числа.

2. Выражение $\{r/(n+1-r)\}\{(1-p)/p\}$ приравнять к значению ЮОа %-ной точки F-распределения с $2(n+1-r_0)$ и $2r_0$ степенями свободы.

3. Найти значение r из полученного уравнения и округлить его «вверх» до ближайшего целого числа, большего r. Если полученное значение r отличается от предыдущей его оценки r_0 более чем на единицу, то шаги 2 и 3 должны быть повторены, причем вместо r_0 следует взять новое значение r.

Аналогично находится значение нижней ЮОа %-ной критической точки. Заметим, что

$$P(X\leqslant r)=P\left(Z\geqslant\frac{n-r}{r+1}\,\frac{p}{1-p}\right), \qquad (10.3.3)$$

где Z есть случайная величина, имеющая F-распределение с $2(r+1)$ и $2(n-r)$ степенями свободы. Поэтому для нахождения нижней ЮОа %-ной точки биномиального распределения может быть применена следующая процедура:

1. Вычислить r_0 — нормальное приближение ЮОа %-ной точки и округлить значение r_0 до ближайшего целого числа.

2. Выражение $\{(n-r)/(r+1)\}\cdot\{p/(1-p)\}$ приравнять к значению ЮОа %-ной точки F-распределения с $2(r_0+1)$ и $2(n-r_0)$ степенями свободы.

3. Найти значение r из полученного уравнения и округлить его «вниз» до ближайшего целого числа, меньшего r. Если полученное значение r отличается от предыдущей его оценки r_0 более чем на единицу, то шаги 2 и 3 должны быть повторены, причем вместо r_0 следует взять новое значение r.

Пример 10.3.1. Найдем верхнюю 5 %-ную точку биномиального (10; 0,3) распределения.

Это распределение имеет среднее значение, равное 3, и дисперсию, равную 2,1. Нормальное приближение верхней 5 %-ной точки равно:

$$r_0=3+1{,}65\times\sqrt{2{,}1}\fallingdotseq 5.$$

Решая уравнение

$$\frac{r}{11-r}\,\frac{0{,}7}{\mathbf{0{,}3}}=2{,}91$$

и производя округление «вверх», получаем значение $r=7$. Это значение отличается от r_0 более чем на единицу; повторим процесс вычислений со значением r_0, равным 7. Тогда получим критическое значение $r=6$, которое может быть подтверждено прямым вычислением или ссылкой на рис. 10.1.1.

Литература: [7, с. 148—150; русский перевод с. 137—140], [47, с. 58—59].

10.4. ПОЛИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Полиномиальное распределение есть естественное многомерное обобщение биномиального распределения[2]. Мы будем описывать

[2] Даже биномиальное распределение может мыслиться как двумерное распределение, но число выпадений решетки полностью определяется знанием числа выпадений герба.

его в терминах подбрасывания k-гранной кости. Если вероятность получить грань i при одном бросании равна p_i $(i = 1, 2, \ldots, k)$ и совершается n независимых бросаний, то вероятность получить n_1 раз первую грань, n_2 раз вторую грань и т. д. равна:

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \cdot p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}, \qquad (10.4.1)$$

где

$$\sum_{i=1}^{k} p_i = 1, \qquad (10.4.2)$$

$$\sum_{i=1}^{k} n_i = n. \qquad (10.4.3)$$

При каждом бросании исходом является либо выпадение г'-й грани с вероятностью p_i, *либо* невыпадение г'-й грани с вероятностью $(1 - p_i)$, т. е. мы имеем дело со схемой Бернулли *. Из этого вытекает, что для числа выпадений г'-й грани справедливо следующее:

$$\text{среднее} = np_i, \qquad (10.4.4)$$

$$\text{дисперсия} = np_i(1 - p_i). \qquad (10.4,5)$$

Число выпадений герба и число выпадений решетки связаны линейным соотношением (их сумма равна *n).* Отсюда соответствующий коэффициент корреляции равен -1, и так как дисперсия каждой из этих величин равна *npq,* то их ковариация должна быть равна $(-npq)$. Ковариация числа выпадений г'-й грани (обозначим его через X_i) и числа выпадений j-й грани (обозначим его через X_j) при n испытаниях равна:

$$\operatorname{cov}(X_i, X_j) = -np_ip_j. \qquad (10.4.6)$$

Коэффициент корреляции равен:

$$\rho_{ij} = \left\{ \frac{p_ip_j}{(1 - p_i)(1 - p_j)} \right\}^{\frac{1}{2}}. \qquad (10.4.7)$$

Методы оценки параметров рассмотрены в параграфе 13.7.

Пример 10.4.1. Правильную шестигранную кость бросают 21 раз. Найдем вероятность того, что при 21 бросании кости получатся следующие результаты:

точно одна «единица», точно четыре «четверки»,
точно две «двойки», точно пять «пятерок»,
точно три «тройки», точно шесть «шестерок».

* Схема Бернулли представляет собой частный случай последовательности независимых испытаний.— *Примеч. пер.*

Вероятность получить такой результат равна:

$$\frac{21!}{1!\,2!\,3!\,4!\,5!\,6!}\left(\frac{1}{6}\right)^{21} = 0{,}000\,094.$$

Литература: [7, с. 206—220; русский перевод с. 193—206], [21, с. 157; русский перевод с. 174—175], [26, с. 48—52], [46, с. 225], [47, с. 281—285], [50, с. 86; русский перевод с. 115—116], [53, с 141; русский перевод с. 198—199], [70, с. 69], [101, с. 138—140; русский перевод с. 151—153].

10.5. АППРОКСИМАЦИЯ ПОЛИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ РАСПРЕДЕЛЕНИЕМ χ^2

Нормальное приближение биномиального распределения обсуждалось в параграфе 10.2. Здесь будет описан более общий результат, применимый к полиномиальному распределению.

Обозначим через *Xi* число выпадений г'-й грани симметричной k-сторонней кости при п бросаниях. Обобщая формулу (10.2.5), можно показать, что случайная величина

$$\sum_{i=1}^{k} (X_i - np_i)^2/(np_i) \qquad (10.5.1)$$

распределена приблизительно как χ^2_{k-1}. Для получения достаточно точных приближений необходимо, чтобы математические ожидания $\{np_i\}$ были не слишком малы. Большинство авторов рекомендуют, чтобы минимальное значение математического ожидания было около 5, хотя У. Кокрен [13] считает, что в качестве минимального его значения можно взять единицу. Формула (10.5.1) может быть записана несколько менее формально, а именно

$$\sum_{\substack{\text{по всем} \\ \text{клеткам}}} (\text{наблюдаемое значение} - \text{математическое ожидание})^2 / \text{математическое ожидание}. \qquad (10.5.2)$$

Уменьшение числа степеней свободы от k до $(k - 1)$ связано с условием, что сумма математических ожиданий наблюдений равна общему числу этих наблюдений. Большое значение величины χ^2 получается, когда некоторые из клеток * содержат значения, заметно отличающиеся от их ожидаемых значений.

Результаты этого параграфа окажутся особенно важными, когда будут изучены таблицы сопряженности признаков (см. параграф 12.9).

Литература: [7, с. 207—210; русский перевод с. 194—198], [45, с. 225—229], [47, с. 285—288], [57, с. 36—38].

* I'M. параграф 12.11.— *Примеч. пер.*

10.6. РАСПРЕДЕЛЕНИЕ ПУАССОНА

Пуассоновское распределение является неотрицательным целочисленным распределением, играющим огромную роль в теории вероятностей и математической статистике. В качестве традиционных примеров случайных величин, подчиняющихся распределению Пуассона, обычно приводят следующие: число альфа-частиц, излучаемых радиоактивным источником за определенный промежуток времени; количество бактерий, видимых под микроскопом на фиксированной части пластинки; число красных кровяных шариков; мутации, вызванные радиацией; число попаданий при бомбардировке Лондона [3]; число прусских солдат, погибших от удара лошади [4]. Распределение дефектов на ткани, помета животных в поле, звезд в пространстве и даже распределение изюминок в пироге также подчиняется закону Пуассона. Это распределение очень часто встречается при изучении проблем, связанных с телефонной сетью. Говорят, что случайная величина X, принимающая целые неотрицательные значения, распределена по закону Пуассона [5], если

$$p(X = r) = e^{-\lambda}\lambda^r/r!, \qquad (10.6.1)$$

где $r = 0, 1, 2, \ldots$ и λ — положительная константа. Нетрудно доказать, что для распределения Пуассона

$$\text{среднее} = Я, \qquad (10.6.2)$$

$$\text{дисперсия} = Я. \qquad (10.6.3)$$

Графики распределений Пуассона со средним 0,9 и 5,0 показаны на рис. 10.6.1. Таблицы распределения Пуассона приводятся в книгах [69] и [76, с. 185—204].

Методы оценки параметров распределения рассмотрены в параграфе 13.8.

Пример 10.6.1. Пекарь выпекает 160 кексов для детского сада, кладя триста изюминок на десять килограммов теста. Какова вероятность того, что какой-нибудь выбранный кекс не будет содержать в себе изюминок?

При решении этой задачи воспользуемся биномиальным распределением. Каждая изюминка с вероятностью, равной 159/160, не попадает в выбранный кекс. Вероятность, что все изюминки не попадут в этот кекс, равна $(159/160)^{300}$, или 0,152.

Чтобы решить эту задачу, можно также применить распределение Пуассона. Среднее число изюминок в каждом кексе равно $300/160 = 1{,}875$. Предположим, что распределение числа изюминок в кексе является пуассоновским со средним значением 1,875.

[3] См. [12].
[4] См. [4].
[5] Напомним, что $\lambda^0 = 1$ и $0! = 1$.

По формуле (10.6.1) получаем:

$$P(\text{изюминки отсутствуют}) = e^{-1{,}875}\,(1{,}875)^0/0! = e^{-1{,}875} = 0{,}153.$$

Результаты при двух способах решения задачи почти совпадают. Связь между этими двумя подходами к решению задачи объясняется в параграфе 10.7.

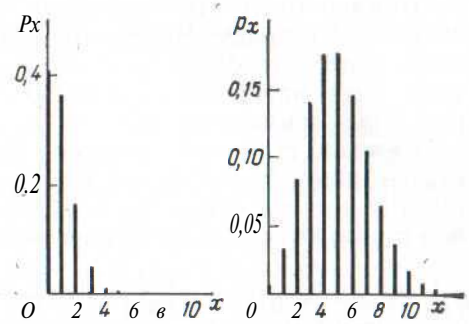


Рис. 10.6.1. Пуассоновские распределения со средними 0,9 и 5,0

Литература: [2, с. 6—20], [3, с. 70—72], [8, с. 303—306], [16, С. 53—54], [21, с. 146—154; русский перевод с. 163—171], [26, С, 45—47], [47, с. 87—121], [53, с. 125—130; русский перевод С. 178—185], [66, с. 32—36], [70, с. 70], [86, с. 119-121], [93, С, 122—140], [102, с. 133—143].

10.7. ПУАССОНОВСКОЕ РАСПРЕДЕЛЕНИЕ КАК ПРЕДЕЛЬНЫЙ СЛУЧАИ БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ

К пуассоновскому распределению можно подойти по-разному. Часто, конечно, исходят из непосредственного его определения, как это было сделано в параграфе 10.6, однако полезно знать, что оно может быть получено также из биномиального распределения.

В качестве примера рассмотрим *большое* число солдат прусской армиии девятнадцатого века и обозначим его через n. Каждый из этих солдат подвергался *очень маленькому* риску умереть от удара лошади, и мы предположим, что вероятность такого случая в данном году для отдельного солдата равна p. Для данного года ожидаемое число смертей от удара лошади определяется биномиальным средним np и его значение будет невелико; обозначим его через Я. Так как p очень мало, q будет очень близко к 1, и биномиальная дисперсия будет очень близка к среднему Я. К предложению об использовании этой ситуации в пуассоновском распределении можно прийти даже интуитивно.

Вероятность того, что в течение рассматриваемого года пройдет r смертей от удара лошади, равна:

$$O''\eta^{n-r},$$

где n — достаточно большое число, p — мало и $np = K$ невелико. Рассмотрим предел этого выражения при $n \to \infty$ и $p \to 0$ (здесь $np = \lambda$, т. е. значение np фиксировано). Предельное значение вероятности равно $e^{-\lambda}\lambda^r/r!$, т. е. вероятность является пуассоновской. В этом смысле пуассоновское распределение иногда считают распределением редких событий.

Пример 10.7.1. Наборщик при работе делает в среднем 1,5 опечатки на страницу. Вычислим вероятность того, что данная страница содержит более чем семь опечаток, а также вероятность того, что в книге объемом в 200 страниц нет страницы, содержащей более семи опечаток.

Каждая страница содержит большое число знаков, и шанс, что выбранный знак будет неверным, очень невелик. Здесь применимо пуассоновское распределение. Вероятность того, что выбранная страница содержит более семи опечаток, равна:

$$1 - e^{-1,5}\left[\frac{(1,5)^0}{0!} + \frac{(1,5)^1}{1!} + \frac{(1,5)^2}{2!} + \cdots + \frac{(1,5)^7}{7!}\right] = 0,000\,170.$$

Вероятность, что выбранная страница содержит семь или менее опечаток, равна 0,999830; вероятность, что никакая из двухсот страниц книги не содержит более семи опечаток, равна $(0,999\,830)^{200} = 0,967$.

Последние результаты можно получить иначе, используя свойства пуассоновского распределения. Будем называть страницу дефектной, если она содержит более семи опечаток. Вероятность, что выбранная страница будет дефектной, очень невелика (0,000 170), но число страниц в книге достаточно велико (200). Среднее число дефектных страниц в книге такого объема равно 200X0,000 170 = = 0,0340. Опираясь на материал параграфа 10.6.1, можно показать, что вероятность отсутствия в книге дефектных страниц равна:

$$e^{-0,0340}\frac{(0,0340)^0}{0!} = 0,967.$$

Пример 10.7.2. Квадрат на рис. 10.7.1 разделен на 900 маленьких квадратиков. Используя таблицы случайных чисел [6], каждый маленький квадратик заполнили точкой с вероятностью $1/25$ и оставили пустым с вероятностью 24/25. Большие, размером 5X5, квадраты, содержащие по 25 маленьких квадратиков, также отмечены на рисунке. (Большой, размером 30x30, квадрат может рассматриваться как участок леса с деревьями, обозначенными точками, или как часть пластинки, которая покрыта бактериями.) Каждый маленький квадратик имеет относительно маленькую вероятность ($p = 1/25$) быть заполненным, но квадрат размером 5x5 состоит из относительно большого числа ($n = 25$) таких маленьких квадратиков. Ожидаемое число точек в расчете на квадрат размером 5x5 равно 1, и в соответствии с нашей теорией вероятность того, что квадрат 5x5 содержит r точек ($r = 0, 1, 2. . .$), приблизительно равна $e^{-1}1^r/r!$ Следовательно, для выбранного квадрата размером 5x5:

P (0 точек) = 0,3679; P (3 точки) = 0,0613;

P (1 точка) = 0,3679; P (4 точки или более) = 0,0190.

P (2 точки) = 0,1839;

Но у нас 36 квадратов размером 5x5. Для совокупности этих 36 квадратов следовало бы ожидать, что $0,3679 \times 36 = 13,2$ квадрата будут пустыми, $0,3679 \times 36 = 13,2$ квадрата будут содержать только по одной точке, $0,1839 \times 36 = 6,7$ квадрата будут содержать по две точки, 0,0613X36 квадратов будут содержать по три точки и 0,0190x36 квадратов будут содержать по четыре или более точек. Эти ожидаемые и действительно наблюдаемые числа приведены в табл. 10.7.1. Там же содержатся аналогичные результаты для 900 маленьких квадратиков. Различия между наблюдениями и ожидаемыми величинами связаны со случайностью выборки.

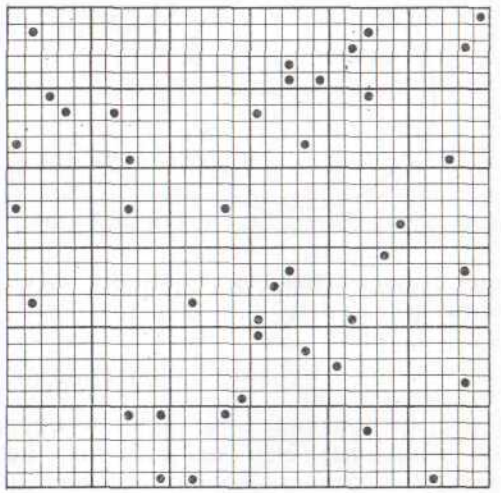


Рис. 10.7.1. Случайные точки на пластинке

Пример 10.7.3. Рассмотрим «бесконечный» лес, в котором деревья распределены случайно, с постоянной плотностью л на единицу площади. Лесничество захотело оценить значение X. В лесу была выбрана случайная точка и измерено расстояние от нее до ближайшего дерева. Если лес засажен плотно, то это расстояние будет, как правило, очень мало, если же лес редкий, то правильным будет обратное утверждение.

Пусть расстояние есть случайная переменная R. Рассмотрим ситуацию, изображенную на рис. 10.7.2. Если ближайшее дерево

[6] См. параграф 14.1.

Таблица 10.7.1. **Пример пуассоновского предела биномиального распределения**

n	Число квадратов, содержащих точно n объектов			
	квадраты размером 5×5		квадраты размером 1×1	
	наблюдаемое	ожидаемое	наблюдаемое	ожидаемое
0	10	13,2	859	864,7
1	15	13,2	41	34,6
2	7	6,7	0	0,7
3	4	2,2	0	0,0
4 и более	0	0,7	0	0,0
Всего	36	36,0	900	900,0

находится на расстоянии r, то круг радиуса r с центром в случайной точке и площадью πr^2 не содержит деревьев. В соответствии с пуассоновским распределением вероятность этого события равна

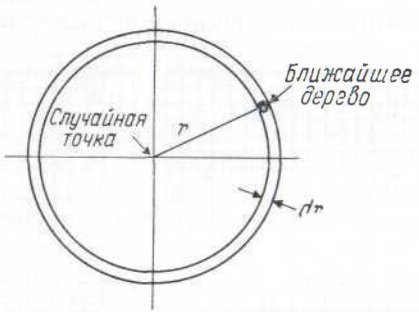


Рис. 10.7.2. Дерево, ближайшее к случайной точке

$\exp(-\pi r^2\lambda)$. С другой стороны, кольцо шириной dr и площадью $2\pi r\,dr$ должно содержать одно дерево, и вероятность этого события равна:

$$\exp(-2\pi r\lambda\,dr)\,|\,(2\pi r\lambda\,dr)^1/1\,! = 2\pi r\lambda\,dr$$

(членами порядка $(dr)^2$ и выше пренебрегаем вследствие их малой величины). Эти два события независимы, и, следовательно, функция плотности вероятностей случайной величины R (расстояния от случайной точки до ближайшего дерева) равна:

$$f(r) = 2\pi r\lambda \exp(-\pi r^2\lambda).$$

Среднее расстояние до ближайшего дерева равно:

$$\int_0^\infty 2\pi r^2\lambda \exp(-\pi r^2\lambda)\,dr = \frac{1}{2}\lambda^{-\frac{1}{2}}.$$

Варианты систем этого типа встречаются в литературе по экологии; приведенный метод может быть применен в задачах по обнаружению дефектов в тканях и других материалах и, более того, распространен на пространство трех измерений, например при изучении звезд в космосе [75, с. 32—33].

Литература: [3, с. 73—77], [21, с. 142—145; русский перевод с. 159—163], [53, с. 125—126; русский перевод с. 178—179], [79, 111—114], [82], [90].

10.8. НОРМАЛЬНОЕ ПРИБЛИЖЕНИЕ ЗАКОНА ПУАССОНА

Вероятность того, что значения случайной величины, распределенной по закону Пуассона со средним Я, заключены между r_1 и r_2 (включительно), равна:

$$\sum_{r=r_1}^{r_2} e^{-\lambda}\lambda^r/r\,!. \qquad (10.8.1)$$

Вычислять значение этой суммы часто очень утомительно. Когда Я больше десяти, центральная предельная теорема[7] позволяет сумму (10.8.1) считать приближенно равной

$$\int_{\alpha_1}^{\alpha_2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx, \qquad (10.8.2)$$

где

$$\alpha_1 = \left(r_1 - \lambda - \frac{1}{2}\right)/\sqrt{\lambda}, \qquad (10.8.3)$$

$$\alpha_2 = \left(r_2 - \lambda + \frac{1}{2}\right)/\sqrt{\lambda}. \qquad (10.8.4)$$

Значение интеграла (10.8.2) легко найти в имеющихся таблицах нормального распределения.

Поправки $\pm 1/2$ в α_1 и α_2 часто опускают, особенно когда значение λ очень велико. Смысл введения этих поправок объясняется (в связи с биномиальным распределением) в параграфе 10.2.

Обозначим результат пуассоновского эксперимента случайной величиной X. Используя (10.8.2), можно утверждать, что $(X-\lambda)/\sqrt{\text{Я}}$ приближенно является единичной нормально распределенной случайной величиной. Из параграфа 9.2 следует, что величина $(X-\lambda)^2/\lambda$ приближенно имеет распределение χ_1^2. Она может быть записана в форме

$$\frac{(\text{наблюдение} - \text{математическое ожидание})^2}{\text{математическое ожидание}}. \qquad (10.8.5)$$

Литература: [47, с. 98—101].

[7] См. параграф 9.1. Сумма некоторого числа независимых пауссоновских случайных величин распределена также по закону Пуассона. Отсюда следует, например, что пуассоновская случайная величина с $\lambda = 10$ может быть представлена как сумма десяти независимых пуассоновских случайных величин с $\lambda = 1$.

10.9. КРИТИЧЕСКИЕ ТОЧКИ ПУАССОНОВСКОГО РАСПРЕДЕЛЕНИЯ

Пусть X — случайная величина, распределенная по закону Пуассона с параметром K. Верхнее ЮОа %-ное критическое значение пуассоновского распределения со средним λ есть целое число r, такое, что

$$\text{p}\,(X \geqslant r) \leqslant \text{a} < \text{P}\,(X \geqslant r - 1). \qquad (10.9.1)$$

Чтобы определить такие критические значения, можно воспользоваться таблицами распределения Пуассона. С помощью нормальной аппроксимации (см. параграф 10.8) можно приближенно вычислить критические точки. Заметим, однако, что

$$P\,(X \geqslant r) = P\,(Y \leqslant 2\lambda), \qquad (10.9.2)$$

где величина $У$ подчиняется распределению χ^2 с 2 степенями свободы [7, с. 173]. Если мы найдем такое r, что значение 2λ лежит между нижними 100α %-ными точками распределения χ^2 с $2r$ и $2(r-1)$ степенями свободы, то r есть верхняя ЮОа %-ная точка пуассоновского распределения со средним л.

Чтобы получить нижнюю 100α %-ную точку, заметим, что

$$P\,(X \leqslant r) = P\,(Z \geqslant 2\lambda), \qquad (10.9.3)$$

где случайная величина Z подчиняется распределению χ^2 с $2(r+1)$ степенями свободы. Если мы найдем такое r, что число 2λ лежит между верхними ЮОа %-ными точками распределения χ^2 с $2r$ и $2(r+1)$ степенями свободы, то r есть нижняя ЮОа %-ная точка пуассоновского распределения со средним Я.

Пример 10.9.1. Найдем верхнюю 5 %-ную точку пуассоновского распределения со средним значением 0,9.

Нужно найти число r, такое, что нижняя 5 %-ная точка распределения χ^2_{2r} больше, чем 1,8, а нижняя 5 %-ная точка распределения $\chi^2_{2(r-1)}$ меньше, чем 1,8. Из таблиц распределений x^2 видно, что $r = 4$; правильность найденного значения можно подтвердить либо прямым вычислением, либо с помощью рис. 10.6.1.

Пример 10.9.2. Найдем нижнюю 10 %-ную точку пуассоновского распределения со средним значением 5,0.

Нужно найти число r, такое, что верхняя 10 %-ная точка распределения χ^2_{2r} меньше 10, а верхняя 10 %-ная точка распределения $\chi^2_{2(r+1)}$ больше 10. Из таблиц распределения χ^2 видно, что $r = 2$ и этот результат может быть подтвержден либо прямым вычислением, либо с помощью рис. 10.6.1.

Литература: [7, с. 173—174; русский перевод с. 163], [47, с. 98—99].

10.10. ГЕОМЕТРИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ (РАСПРЕДЕЛЕНИЕПАСКАЛЯ)

Аналогично пауссоновскому геометрическое распределение есть распределение случайной величины, принимающей неотрицательные целочисленные значения. Мы можем к нему прийти, рассмотрев эксперимент, в ходе которого человек подбрасывает монету до тех пор, пока не выпадет герб. Если вероятность выпадения герба при одном подбрасывании равна p и $q = 1 - p$, то вероятность выпадения решетки j раз, прежде чем выпадет первый герб, равна:

$$p_j = q^j p \quad (j = 0,\ 1,\ 2,\ \ldots). \qquad (10.10.1)$$

Эти вероятности определяют *геометрическое распределение.* Математическое ожидание (среднее) и дисперсия его следующие:

$$\text{среднее} = q/p, \qquad (10.10.2)$$

$$\text{дисперсия} = q/p^2. \qquad (10.10.3)$$

График геометрического распределения с $p = 0{,}3$ показан на рис. 10.11.1. Надо отметить, что геометрическое распределение есть частный случай отрицательного биномиального распределения.

Методы оценки параметров рассмотрены в параграфе 13.9.

Литература: [47, с. 123—124].

10.11. ОТРИЦАТЕЛЬНОЕ БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Отрицательное биномиальное распределение есть также распределение на множестве неотрицательных целых чисел; оно определяется с помощью двух параметров — k и p. Мы, как обычно, полагаем, что $q = 1 - p$. Это распределение фактически является распределением суммы k взаимно независимых, геометрически распределенных случайных величин или же распределением числа решеток, выпавших перед k-м выпадением герба в эксперименте с монетой; оно имеет следующий вид:

$$p_j = \binom{k+j-1}{j} p^k q^j \quad (j = 0,\ 1,\ 2,\ \ldots). \qquad (10.11.1)$$

Комбинаторные обозначения объясняются в параграфе 1.5.

Параметрами отрицательного биномиального распределения являются p и k. Значения параметра p должны лежать между О и 1. Значения параметра k при выводе формулы (10.11.1) были ограничены целыми положительными числами. Но фактически можно предположить, что k может принимать любые положительные значения, так как формула (10.11.1) определяет распределение для всех положительных k. Среднее значение и дисперсия распределения равны:

$$\text{среднее} = \frac{kq}{p}, \qquad (10.11.2)$$

$$\text{дисперсия} = \frac{kq}{p^2}. \qquad (10.11.3)$$

Два примера графиков отрицательного биномиального распределения приведены на рис. 10.11.1. Когда $k = 1$, это распределение становится геометрическим распределением.

Случайная величина, имеющая отрицательное биномиальное распределение с параметрами p и k, может быть представлена как сумма k независимых случайных величин, имеющих геометрическое распределение с параметром p. Когда значение k велико, центральная предельная теорема позволяет нам использовать следующее приближение для частичной суммы «последовательных» вероятностей отрицательного биномиального распределения:

$$\sum_{j=r_1}^{r_2} p_j \doteq \int_{\alpha_1}^{\alpha_2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx, \qquad (10.11.4)$$

где

$$\alpha_1 = \left(r_1 - kq/p - \frac{1}{2}\right)/(kq/p^2)^{\frac{1}{2}}, \qquad (10.11.5)$$

$$\alpha_2 = \left(r_2 - kq/p + \frac{1}{2}\right)/(kq/p^2)^{\frac{1}{2}}. \qquad (10.11.6)$$

Поправки $\pm 1/2$ объяснены (в связи с биномиальным распределением) в параграфе 10.2. Ими часто пренебрегают, особенно когда k очень велико.

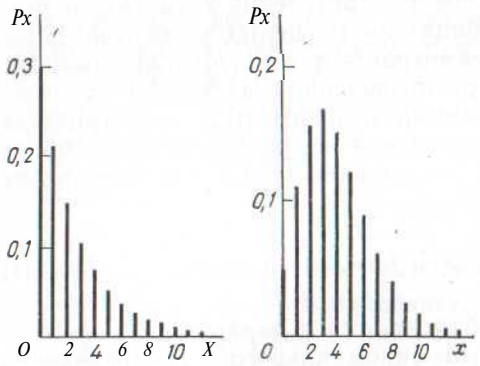


Рис. 10.11.1. Отрицательные биномиальные распределения с параметрами $k=1$, $p=0,3$ и $k=6$, $p=0,6$ соответственно. Первое распределение называется также геометрическим распределением

Из материала параграфа 10.6 известно, что математическое ожидание и дисперсия пуассоновского распределения равны друг другу. С другой стороны, дисперсия биномиального распределения меньше, чем среднее (см. параграф 10.1), и в этом смысле биномиальное распределение иногда называют *малодисперсным*. Очевидно, что дисперсия отрицательного биномиального распределения больше, чем среднее. Говорят, что это распределение *сверхдисперсно*. Это свойство часто используется в тех случаях, когда необходимо подобрать целочисленное неотрицательное распределение к данным с большой дисперсией. Дисперсия и среднее отрицательного биномиального распределения, имеющего большое k и умеренное среднее, приблизительно равны (так как значение параметра p ограничено единицей). Фактически это позволяет доказать, что предельное распределение есть распределение Пуассона.

Таблицы отрицательного биномиального распределения приведены в книге Уильямсона и Бретертона [104]. Методы оценки параметров рассматриваются в параграфе 13.10.

Пример 10.11.1. Подберем дискретное распределение к частотным данным табл. 10.11.1.

Среднее число яиц равно 6,73. Распределение Пуассона со средним значением, равным 6,73, имеет модальное значение $j = 6$. Данное распределение в точности соответствует этим данным. Попытаемся подобрать отрицательное биномиальное распределение методом моментов (см. параграф 13.3). Приравняем среднее и дисперсию отрицательного биномиального распределения к соответствующим выборочным значениям. Таким образом,

$$kq/p = 6,73 \text{ и } kq/p^2 = 116,195.$$

Разделив первое уравнение на второе, получим $p = 0,057\,948\,286$, откуда следует, что $q = 1 - p = 0,942051\,714$ и $k = 0,414\,186524$. Так как

$$\binom{k-1}{0} = 1 \text{ и } p_{j+1} = \{(k+j)\,q/(j+1)\}\,p_j,$$

подбираемые частоты $\{f_j\}$ могут быть вычислены следующим образом:

$$f_0 = 90 \times (0,057\,948\,286)^{0,414\,186\,524}(0,942\,051\,714)^0 = 27,7;$$

$$f_1 = f_0 \times (0,414\,186\,524) \times (0,942\,051\,714)/1 = 10,8;$$

$$f_2 = /1 \times (1,414\,186\,524) \times (0,942\,051\,714)/2 = 7,2;$$

$$f_3 = f_2 \times (2,414\,186\,524) \times (0,942\,051\,714)/3 = 5,5.$$

Подобранная частота f_0 очень велика, в то же время значения других частот гораздо меньше. Отрицательное биномиальное распределение с меньшим значением f_0 может быть получено с помощью того же среднего значения 6,73, но с большим значением k (и соответственно меньшей дисперсией). Такие распределения представлены в табл. 10.11.1. Из приведенных в таблице распределений лучшее согласие с данными наблюдений дает отрицательное биномиальное распределение при $k = 0,6$ [8].

Отметим сходство между пуассоновским распределением со средним 6,73 и отрицательным биномиальным распределением с тем же средним и $k = 100$. Оба этих распределения не подходят к данным наблюдений.

[8] Для проверки адекватности подобранного распределения может быть использован критерий согласия χ^2 (см. параграф 12.11) или критерий Колмогорова—Смирнова (см. параграф 12.12).

Таблица 10.11.1. Данные о распределении яиц паразитических нематод у шотландских овец и соответствующие значения частот, вычисленные на основе биномиального и пуассоновского распределений

Число яиц	Наблюдаемые частоты	Подобранные частоты: отрицательное биномиальное распределение со средним 6,73 и параметром k, равным 0,414 19	0,5	0,6	100	Пуассоновское распределение со средним и S 6,73
0	20	27,7	23,7	20,0	0,1	0,1
1	12	10,8	11,0	11,0	0,8	0,7
2	14	7,2	7,7	8,1	2,7	2,4
3	7	5,5	6,0	6,5	5,7	5,5
4	3	4,4	4,9	5,3	9,3	9,2
5	6	3,6	4,1	4,5	12,2	12,4
6	3	3,1	3,5	3,9	13,5	13,9
7	3	2,7	3,0	3,3	12,9	13,3
8	2	2,3	2,6	2,9	10,9	11,2
9	3	2,1	2,3	2,6	8,2	8,4
10—14	4	7,4	8,2	9,0	13,0	12,5
15-19	2	4,5	4,8	5,1	0,5	0,4
20+	11	8,7	8,2	7,8	0,2	0,0
Всего	90	90,0	90,0	90,0	90,0	90,0

Источник. [103, табл. 119].

Литература: [21, с. 155—156; русский перевод с. 171—174], [26, с. 55—58], *[47,* с. 122—142], [53, с. 130—131; русский перевод с. 185—186], [103, с. 5—16].

10.12. ФИШЕРОВСКОЕ РАСПРЕДЕЛЕНИЕ ПО ЛОГАРИФМИЧЕСКОМУ РЯДУ

Р. А. Фишер в сотрудничестве с А. С. Корбетом и С. Б. Уильямсом развил математическую теорию, которая с некоторым успехом применяется при описании относительных частот появления различных биологических видов в случайных выборках из неоднородных популяций. Уильяме показал, что логарифмический ряд может быть применен к большому числу биологических проблем, в которых целое n может означать число разновидностей в некотором виде, число видов в подсемействе, число паразитов в организме-«хозяине» и даже число научных работ, опубликованных в некотором году, в расчете на одного биолога. Трудно поверить, что когда-либо будет создана единая теория, объясняющая уместность использования логарифмического ряда во всех этих задачах, но тем не менее это распределение часто применяется в биологии.

Распределение задано на множестве целых положительных чисел (исключается нуль):

$$P(X = j) = \{-\ln(1-\alpha)\}^{-1}\alpha^j/j. \qquad (10.12.1)$$

Параметр a лежит строго между нулем и единицей, j принимает значения 1, 2, 3, ... Отметим, что распределение имеет

$$\text{среднее} = \{-\ln(1-\alpha)\}^{-1}\alpha/(1-\alpha), \qquad (10.12.2)$$

$$\text{дисперсия} = \frac{\text{среднее}}{1-\alpha} - (\text{среднее})^2.$$

Данное распределение может быть получено из отрицательного биномиального распределения. Когда параметр k отрицательного биномиального распределения очень мал, условное распределение отрицательной биномиальной случайной величины (при условии, что она не равна нулю) становится распределением по логарифмическому ряду.

Методы оценки параметров рассмотрены в параграфе 13.11.

Пример 10.12.1. Биолог использует световую ловушку, чтобы поймать большое количество мотыльков. Он знает, что существует очень много различных видов мотыльков, но фактическое число их ему не известно. Он знает также, что вероятность поймать мотылька какого-либо вида в течение определенного промежутка времени есть величина постоянная и независящая от его вида. Биолог продолжает ловить, пока не поймает n мотыльков, где n весьма велико. Какое распределение числа мотыльков различных видов он мог бы ожидать?

Биолог продолжает ловить мотыльков, пока не получит некоторое определенное общее число их. Число видов очень велико, так что какой-либо определенный вид может не, быть представлен во всей совокупности пойманных мотыльков с большой вероятностью. В сущности, биолог производит отбор образцов до тех пор, пока не получит в среднем по k мотыльков каждого вида, где k меньше единицы и весьма мало. Итак, для каждого вида число пойманных мотыльков следует отрицательному биномиальному распределению, в котором параметр k мал. Мы делаем вывод, что для каждого вида, представленного в выборке, число пойманных мотыльков является случайным наблюдением из совокупности, распределенной по фишеровскому логарифмическому ряду. Предполагаем, что все виды одинаково представлены в генеральной совокупности и имеют одинаковые вероятности поимки. Следовательно, мы делаем вывод, что число видов мотыльков, представленных пойманными особями, следует распределению по логарифмическому ряду.

На практике многие из этих предположений не могут быть строго соблюдены. Несмотря на это, С. Б. Уильяме показал, что распределение по логарифмическому ряду дает хорошее согласие с данными наблюдений при многих различных условиях.

Может показаться, что, выполнив такой эксперимент, мы можем вернуться к отрицательному биномиальному распределению, чтобы определить общее число видов, включая и те, которые не представлены пойманными особями. Но это невозможно, так как наши предположения включают достаточно сильное условие о бесконечности общего числа видов.

Пример 10.12.2. По табл. 10.12.1 подберем распределение по логарифмическому ряду к данным о мотыльках. Среднее число особей, представляющих вид, равно 6814/197 — 34,588 832, и мы приравниваем (10.12.2) к этому значению. Методом проб и ошибок находим, что a = 0,994 47; частоты подобранного распределения в табл. 10.12.1 вычисляются следующим образом:

Таблица 10.12.1. **Распределение по видам числа особей липидоптеры, пойманных световой ловушкой на Ротамстедской экспериментальной биостанции в 1935 г. Общее число пойманных особей равно 6814.**

Число особей, представляющих вид, j	Число видов, представленных особями	
	зафиксированное при наблюдении	соответствующее подобранному распределению
1	37	37,7
2	22	18,7
3	12	12,4
4	12	9,3
5	11	7,4
6	11	6,1
7	6	5,2
8	4	4,5
9	3	4,0
10	5	3,6
П	2	3,2
12	4	3,0
13	2	2,7
14	3	2,5
15	2	2,3
16+	61	74,4
Всего	197	197,0

Источник. [103, с. 25, табл. 8].

$$f_1 = 197 \times 0{,}994\,47/\{-1\text{п}\,(0{,}005\,53)\} = 37{,}7;$$
$$f_2 = f_1 \times 0{,}994\,47 \times 1/2 = 18{,}7;$$
$$f_3 = f_2 \times 0{,}994\,47 \times 2/3 = 12{,}4;$$
$$f_4 = f_3 \times 0{,}99447 \times 3/4 = 9{,}3.$$

Критерий согласия χ^2 из параграфа 12.11 (или критерий Колмогорова—Смирнова, рассмотренный в параграфе 12.12) может быть применен для проверки адекватности подбора.

Литература: [47, с. 166—182], [53, с. 131—133; русский перевод с. 186—188], [103, с. 5—16].

10.13. ГИПЕРГЕОМЕТРИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ

Рассмотрим совокупность, содержащую R красных шаров и B черных шаров. Общий объем совокупности равен $R+B$. Из этой совокупности случайным образом выбираются n шаров (безвозвратная выборка). Распределение числа красных шаров в выборке известно как *гипергеометрическое распределение*:

$$P(r \text{ красных шаров}) = \binom{R}{r}\binom{B}{n-r}\Big/\binom{R+B}{n}. \qquad (10.13.1)$$

Объем выборки n должен быть меньше, чем $R+B$ и $0 < r < \min(R, n)$. Отметим, что

$$\text{среднее} = nR/(R+B), \qquad (10.13.2)$$

$$\text{дисперсия} = \frac{nRB}{(R+B)^2}\left(1 + \frac{n-1}{R+B-1}\right) \qquad (10.13.3)$$

Когда объем совокупности велик, а объем выборки сравнительно мал, данное распределение является приближенно биномиальным с параметрами n и $p = R/(R+B)$. Далее, когда n велико, $(R+B)$ очень велико, p, равное $B/(R+B)$, мало, а np принимает умеренное значение, распределение является приближенно пуассоновским (см. параграф 10.7). При некоторых условиях в качестве приближения может быть также использовано нормальное распределение (см. параграфы 10.2 и 10.8). Таблицы гипергеометрического распределения составлены Либерманом и Оуэном [59].

Пирсоновская система частотных кривых имеет свое начало в гипергеометрическом распределении.

Литература: [21, с. 41—44; русский перевод с. 55—59], [26, с. 53—55], [47, с. 143—165], [50, с. 82—85; русский перевод с. 114—115], [53, с. 133—135; русский перевод с. 188—191].

10.14. УПРАЖНЕНИЯ

1. Табл. 10.14.1 содержит некоторые итоговые данные по заражению паразитами. Вычислите среднее и дисперсию распределения. Выборочная дисперсия меньше, чем среднее. Подберите к этим данным биномиальное распределение, считая np и npq выборочным средним и выборочной дисперсией соответственно.

Таблица 10.14.1. **Итоги подсчета случаев заражения паразитами**

Количество паразитов	Число животных	Количество паразитов	Число животных
Меньше 8	0	15	6
8	2	16	3
9	0	17	1
10	2	18	3
11	2	Больше 18	0
12	5		
13	2	Всего	30
14	4		

2. Подберите к данным табл. 10.14.1 пуассоновское распределение, считая λ выборочным средним.

3. Симметричную шестигранную игральную кость бросают 1000 раз. Вычислите вероятность того, что 200 или более раз выпадет «шестерка».

Указание. При каждом бросании вероятность получить «шестерку» равна $^1/_6$; «шестерка» не выпадает с вероятностью $^5/_6$.

4. Симметричную шестигранную игральную кость бросают 10 раз. Вычислите вероятность того, что точно 4 раза выпадет «шестерка» и 2 раза выпадет «тройка».

Указание. При каждом бросании «шестерка» и «тройка» выпадают с вероятностями, каждая из которых равна $^1/_6$; вероятность получить что-либо еще равна $^2/_3$.

5. Южная часть Лондона во время второй мировой войны была разделена на 576 небольших квадратов одинакового размера. Р. Д. Кларк [12] в своем докладе привел следующие данные: 229 квадратов не были поражены бомбовыми ударами, 211 квадратов получили точно одно попадание, по два попадания было в 93 квадратах, 35 квадратов получили три попадания, 7 квадратов — четыре попадания и на один квадрат пришлось не менее пяти бомбовых ударов. Подберите пуассоновское распределение к этим данным.

Указание. Приравняйте пуассоновское среднее (10.6.2) к среднему наблюдаемого распределения.

6. Приравняйте среднее геометрического распределения (10.10.2) и дисперсию (10.10.3) к выборочным значениям этих моментов в числовом примере 10.11.1. Затем вычислите частоты подобранного геометрического распределения, соответствующие данным табл. 10.11.1.

7. Коробка содержит 10 000 красных шаров и 20 000 черных шаров; случайным образом из коробки выбирается 100 шаров (безвозвратная выборка). Найдите вероятность того, что 40 или более выбранных шаров являются красными.

8. С помощью метода из параграфа 10.3 найдите нижнюю и верхнюю 2,5 %-ные точки биномиального распределения с параметрами $n=18$, $p=0{,}4$.

9. С помощью метода из параграфа 10.9 найдите нижнюю и верхнюю 5 %-ные точки пуассоновского распределения со средним $\lambda=1{,}5$.

И. СИСТЕМА ФУНКЦИЙ ПЛОТНОСТИ ПИРСОНА

Разработанная Пирсоном система функций плотности вероятностей определяется с помощью дифференциального уравнения, которое по форме сходно с разностным уравнением гипергеометрического распределения. Изменением параметров данного уравнения можно добиться нужного изменения в значениях показателей асимметрии и эксцесса полученных непрерывных распределений. Значения этих параметров полностью определены первыми четырьмя моментами распределения, соответственно для подбора кривых используются значения первых четырех выборочных моментов. Данный метод удобен в тех случаях, когда необходимо просто подобрать подходящую кривую плотности и не стоит задача обоснования типа функциональной зависимости, соответствующей такой кривой.

11.1. ВВЕДЕНИЕ

Меняя параметры R, B и n, задающие гипергеометрическое распределение, можно получить много различных типов дискретных распределений (см. параграф 10.13). Например, в случае когда $R = B$ и оба эти значения очень велики, при достаточно большом объеме выборки (п) распределение числа красных шаров в выборке несущественно отличается от биномиального с параметрами

// (большое значение) и $p = \frac{1}{2}$. Далее можно получить симметричное распределение, близкое к нормальному. Изменением параметров гипергеометрического распределения можно добиться нужного изменения в значениях показателей асимметрии и эксцесса получаемых распределений (см. параграф 8.6). Карл Пирсон заметил этот факт и разработал пирсоновскую систему функций плотности вероятностей, используя дифференциальное уравнение, сходное с разностным уравнением дискретного гипергеометрического распределения.

В процессе подбора пирсоновской функции плотности вероятностей необходимо выполнить очень много элементарных арифметических операций, однако сейчас с появлением настольных калькуляторов, автоматически вычисляющих значения экспоненциальных, логарифмических и тригонометрических функций, такого рода работа требует усилий неизмеримо меньше, чем это было при счете вручную. Этот метод весьма удобен в тех случаях, когда нужно просто подобрать подходящую кривую плотности и нет необходимости обосновать тип функциональной зависимости, соответствующей такой кривой.

При описании гипергеометрического распределения была приведена формула (10.13.1). Обозначим вероятность в левой части этой формулы символом p_r. Нетрудно показать, что разность $\Delta p_r = p_{r+1} - p_r$ удовлетворяет разностному уравнению

$$\frac{1}{p_r}\Delta p_r = -\left\{\frac{-(nR - B + n - 1) + (R + B + 2)\,r}{(B - n + 1) + (B - n + 2)\,r + r^2}\right\}\Gamma \qquad (11.1.1)$$

В рамках системы Пирсона функция плотности вероятностей удовлетворяет дифференциальному уравнению вида

$$\frac{1}{f(x)}\frac{df}{dx} = -\frac{a + x}{c_0 + c_1 x + c_2 x^2}, \qquad (11.1.2)$$

где a, c_0, c_1 и c_2 — соответствующие константы.

Литература: [9, с. 155—159], [20, с. 35—46], [48, с. 9—14].

11.2. ПОДБОР КРИВОЙ ПИРСОНА

Дифференциальное уравнение (11.1.2) определяет пирсоновскую систему кривых, значения констант a, c_0, c_1 и c_2 однозначно заданы значениями первых четырех моментов распределения. Для того чтобы подобрать кривую Пирсона, которая соответствовала бы некоторой совокупности данных, применяют метод моментов (см. параграф 13.3). Вычисляют значения первых четырех начальных моментов m'_1, m'_2, m'_3 и m'_4, центральных моментов m_2, m_3 и m_4, а также функций моментов $\sqrt{b_1}$ и b_2 (см. параграф 8.6). Если начало координат выбрано в точке, соответствующей среднему, и если величина d определена формулой

$$d = 2(5b_2 - 6b_1 - 9), \qquad (11.2.1)$$

то значения констант дифференциального уравнения задаются выражениями

$$в = \{\sqrt{m_2}(b_2 + 3)\sqrt{b_1}\}/d, \qquad (11.2.2)$$

$$c_0 = m_2(4b_2 - 3b_1)/d, \qquad (11.2.3)$$

$$c_1 = a, \qquad (11.2.4)$$

$$c_2 = (2b_2 - 3b_1 - 6)/d. \qquad (11.2.5)$$

Тип полученной кривой зависит от значений корней квадратного уравнения, соответствующего квадратному трехчлену, в знаменателе правой части (11.1.2). Пирсон выявил двенадцать различных типов кривых (они кратко описаны, например, в работе У. П. Элдертона и Н. Л. Джонсона [20, с. 45]). Основные типы имеют номера I, IV и VI. В пограничных ситуациях, когда один из основных типов переходит в другой, получаются более простые формы кривых перехода. Например, кривая нормального распределения получается при $c_1 = cг = 0$ ($\sqrt{b_1} = 0$; $b_2 = 3$).

Во многих работах подробно описаны процедуры ручного счета при подборе кривых различного типа (см., например, [20]). В этих процедурах используются громоздкие формулы, скрывающие подлинную простоту метода. В приводимых далее примерах мы подберем кривые, основываясь на исходных принципах метода и не будем применять эти формулы. Следует обратить внимание на критерий Пирсона

$$k = c_1^2/(Ac_0) \qquad (11.2.6)$$

выделяющий различные типы кривых и основанный на различении типов корней квадратного уравнения в знаменателе правой части формулы (11.1.2).

Когда кривая плотности вероятностей уже выбрана, для оценки адекватности подбора можно воспользоваться критерием согласия χ^2, описанным в параграфе 12.11 (или, возможно, критерием Колмогорова—Смирнова, описанным в параграфе 12.12).

Метод Пирсона можно применять для подбора кривых распределения к таким частотным данным, которые нельзя рассматривать как результаты независимых наблюдений значений некоторой одномерной случайной величины (например, такие данные могут характеризовать распределение населения некоторой страны по возрастным группам); в таких случаях нельзя применять критерий согласия Колмогорова—Смирнова или χ^2.

Пример **11.2.1.** Подберем кривую Пирсона к данным, приведенным в табл. 11.2.1. Процедура начинается с вычисления значений моментов и соответствующих функций моментов:

$$m_1' = 37{,}875;$$

$$m_2' = 1626{,}075; \quad m_2 = 191{,}559\,375; \quad \sqrt{b_1} = 0{,}712246085;$$

$$m_3' = 77\,986{,}575; \quad m_3 = 1888{,}36172; \quad b_1 = 0{,}507294\,486;$$

$$m_4' = 4\,100\,394{,}675; \quad m_4 = 107703{,}2971; \quad b_2 = 2{,}935\,095\,088.$$

Таблица 11.2.1. **Некоторые данные из работы У. П. Элдертона. Приведены также соответствующие значения для сглаживающей кривой**

Центральное значение возрастной группы	Наблюдавшаяся частота	Наблюдавшаяся частота, деленная на 5	Сглаженное значение в центре интервала X 1000
17	34	6,8	6,7
22	145	29,0	27,8
27	156	31,2	30,1
32	145	29,0	28,6
37	123	24,6	25,4
42	103	20,6	21,6
47	86	17,2	17,6
52	71	14,2	13,7
57	55	11,0	10,2
62	37	7,4	7,2
67	21	4,2	4,7
72	13	2,6	2,8
77	7	1,4	1,4
82	3	0,6	0,6
87	1	0,2	0,1
	Всего 1000	—	—

Источник. [20, с. 53].

Формулы для вычислений приведены в параграфе 8.6. Далее мы воспользуемся формулами с (11.2.2) по (11.2.5) для вычисления значений констант:

$$a = 11{,}115\,824\,33; \quad c_1 = 11{,}115\,82433;$$

$$c_0 = 371{,}896\,975; \quad c_2 = -0{,}313\,806\,272.$$

Итак, дифференциальное уравнение (11.1.2) в данном случае имеет вид

$$\frac{1}{f}\frac{df}{dx} = \frac{11{,}11582433 + x}{371{,}893975 + 11{,}115\,824\,33x - 0{,}313\,806\,272x^2}.$$

Корни квадратного уравнения, соответствующего трехчлену знаменателя, вещественны, их значения равны $-21{,}003\,132\,03$ и $56{,}425\,701\,06$. Отсюда следует, что правую часть дифференциального уравнения можно записать в виде

$$\frac{A}{21{,}003\,132\,03 + x} + \frac{B}{56{,}425701\,06 - x}.$$

Значения констант Л и В можно найти, решив систему из двух уравнений. Каждое из этих уравнений может быть получено приравниванием двух выражений для правой части дифференциального уравнения. Для получения двух уравнений необходимо взять /ми произвольных значения переменной x и подставить их пооче-

редно в указанное равенство. Например, при использовании значений $x = 0$ и $x = 10$ мы приходим к системе

$$0{,}047\,611\,946A + 0{,}017\,722\,420B = -0{,}029\,889\,526;$$
$$0{,}032\,254\,805A + 0{,}021\,539\,793B = -0{,}046\,750\,082,$$

откуда $A = 0{,}406\,924\,333$ и $B = -2{,}779\,754\,990$. Правильность этого решения можно проверить, подставив в рассмотренное равенство какое-либо третье значение переменной x, например, $x = 100$. При этом для обеих частей равенства получаются соответственно значения 0,067 156 379 и 0,067 156 376, которые близки друг к другу.

Левая часть дифференциального уравнения является производной функции $\ln f$ по переменной x, следовательно, можно записать:

$$\frac{d}{dx}\ln f = \frac{0{,}406\,924\,333}{21{,}003\,132\,03 + x} - \frac{2{,}779\,754\,990}{56{,}425\,701\,06 - x}.$$

Поскольку $\frac{d}{dx}\ln(\alpha + \beta x) = \beta/(\alpha + \beta x)$, справедливо соотношение

$$\ln f = 0{,}406\,924\,333\ln(21{,}003\,132\,03 + x) + 2{,}779\,754\,990\ln(56{,}425\,701\,06 - x) + const$$

и $f = f_0(\gamma + x)^u(\delta - x)^v$, где

$$\gamma = 21{,}003\,132\,03; \quad u = 0{,}406\,924\,333;$$
$$\delta = 56{,}425\,701\,06; \quad v = 2{,}779\,754\,990.$$

Данная кривая относится к типу I по Пирсону и является графиком функции плотности бета-распределения случайной величины x, которая принимает значения в диапазоне $-\gamma \leqslant x \leqslant \delta$.

Для того чтобы площадь под выбранной кривой была равна единице, необходимо взять следующее значение константы f_0[1]:

$$f_0 = \left\{(\gamma + \delta)^{u+v+1}\frac{\Gamma(u+1)\Gamma(v+1)}{\Gamma(u+v+2)}\right\}^{-1}.$$

Вычисляя, находим, что $f_0 = 0{,}970\,08 \times 10^{-7}$. При возвращении к исходной системе координат (от системы, начало которой было смещено в точку среднего значения) частотная функция приобретает вид

$$f_0'(y - m_1' + x)^u(\delta + m_1' - x)^v,$$

величина x в исходной системе координат принимает значения в диапазоне от $m_1' - \gamma$ до $m_1' + \delta$. Подобранная кривая и исходные данные сопоставлены в табл. 11.2.1 и на рис. 11.2.1. Площадь участков, лежащих под теми или иными отрезками кривой, может быть найдена методами приближенного интегрирования (см. гл. 5).

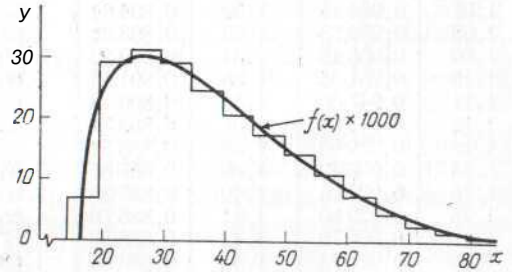


Рис. 11.2.1. Применение пирсоновской кривой типа I для сглаживания некоторого набора частотных данных из работы У. П. Элдертона

Данные, приведенные в табл. 11.2.1, не являются независимыми наблюдениями значений какой-либо одномерной случайной величины, поэтому критерии согласия, описанные в параграфах 12.11 и 12.12, в данном случае не могут применяться.

Пример 11.2.2. Подберем кривую Пирсона к данным, приведенным в табл. 11.2.3. Как и в предыдущем случае, процедура начинается с вычисления значений первых четырех моментов и связанных с ними величин. Получаем

$$m_1' = 44{,}577\,233\,99;$$
$$m_2' = 2102{,}403\,320; \quad m_2 = 115{,}273\,530\,2; \quad \sqrt{b_1} = -0{,}071\,273462;$$
$$m_3' = 10398{,}908{,}2641; \quad m_3 = -88{,}210\,900; \quad b_1 = 0{,}005\,079\,906;$$
$$m_4' = 5349410{,}913; \quad m_4 = 42\,074{,}106\,00; \quad b_2 = 3{,}166\,326\,603;$$
$$a = -0{,}346\,901\,821; \quad c_1 = -0{,}346\,901\,821;$$
$$c_0 = 107{,}2037139; \quad c_2 = 0{,}023\,335\,268.$$

И данном случае корни квадратного уравнения, соответствующего знаменателю правой части дифференциального уравнения, комплексны. Поэтому квадратный трехчлен мы перегруппируем следующим образом:

$$\text{знаменатель} = 0{,}023\,335\,268x^2 - 0{,}346\,901\,821x + 107{,}203\,7139 =$$
$$= 0{,}023\,335\,268\,(x^2 - 14{,}865\,988\,29x + 4\,594{,}063\,967) =$$
$$= 0{,}023\,335\,268\,\{(x - 7{,}432\,994\,15)^2 + 4\,538{,}814\,565\} =$$
$$= c_2\{(x - \alpha)^2 + \beta^2\}.$$

[1] Численные значения гамма-функции $\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx$ приведены в табл. 11.2.2. При $n > 2$ для определения значений функции используют соотношение $\Gamma(n) = (n-1)\,\Gamma(n-1)$. (Заметим, что $\Gamma(n) = (n-1)!$, если n — положительное целое число). Для больших положительных значений n очень точное приближение задается формулой Стирлинга

$$\Gamma(n) \cong \text{гс"е-«}\ (2\text{я}/\text{л})^{\frac{1}{2}}.$$

Таблица 11.2.2. **Гамма-функция $\Gamma(n)$**

n	Г (л)	n	Г (я)	n	Г (n)	n	Г (в)
1,00	1,000 00	1,25	0,906 40	1,50	0,886 23	1,75	0,91906
1,01	0,994 33	1,26	0,904 40	1,51	0,886 59	1,76	0,921 37
1,02	0,988 84	1,27	0,902 50	1,52	0,887 04	1,77	0,923 76
1,03	0,983 55	1,28	0,900 72	1,53	0,887 57	1,78	0,926 23
1,04	0,978 44	1,29	0,899 04	1,54	0,888 18	1,79	0,928 77
1,05	0,973 50	1,30	0,897 47	1,55	0,888 87	1,80	0,931 38
1,06	0,968 74	1,31	0,896 00	1,56	0,889 64	1,81	0,934 08
1,07	0,964 15	1,32	0,894 64	1,57	0,890 49	1,82	0,936 85
1,08	0,959 73	1,33	0,893 38	1,58	0,891 42	1,83	0,939 69
1,09	0,955 46	1,34	0,892 22	1,59	0,892 43	1,84	0,942 61
1,10	0,951 35	1,35	0,891 15	1,60	0,893 52	1,85	0,94561
1,11	0,947 39	1,36	0,890 18	1,61	0,894 68	1,86	0,948 69
1,12	0,943 59	1,37	0,88931	1,62	0,895 92	1,87	0,951 84
1,13	0,939 93	1,38	0,888 54	1,63	0,897 24	1,88	0,955 07
1,14	0,936 42	1,39	0,88785	1,64	0,898 64	1,89	0,958 38
1,15	0,933 04	1,40	0,887 26	1,65	0,900 12	1,90	0,961 77
1,16	0,929 80	1,41	0,886 76	1,66	0,901 67	1,91	0,965 23
1,17	0,926 70	1,42	0,886 36	1,67	0,903 30	1,92	0,968 78
1,18	0,923 73	1,43	0,886 04	1,68	0,905 00	1,93	0,972 40
1,19	0,920 88	1,44	0,885 80	1,69	0,906 78	1,94	0,976 10
1,20	0,918 17	1,45	0,885 65	1,70	0,908 64	1,95	0,979 88
1,21	0,91558	1,46	0,885 60	1,71	0,91057	1,96	0,983 74
1,22	0,913 11	1,47	0,885 63	1,72	0,91258	1,97	0,987 68
1,23	0,91075	1,48	0,885 75	1,73	0,91466	1,98	0,991 71
1,24	0,908 52	1,49	0,885 95	1,74	0,91683	1,99	0,99581
						2,00	1,000 00

Источник. Standard Mathematical Tables (Twelth Edition), C. D. Hodgman (ed.). The Chemical Rubber Co., 1963. (Публикация данной таблицы разрешена фирмой The Chemical Rubber Co.).

Перегруппируем также и числитель:

$$\text{числитель} = x - 0{,}346\,901\,821 =$$
$$= x \_ 7{,}432\,994\,15 + 7{,}086\,092\,329 =$$
$$= x - a + \mathrm{Y}.$$

Теперь мы можем решить данное дифференциальное уравнение, которое удобно переписать в виде

$$\frac{d}{Ax}\ln f = \frac{-2(x-\alpha)}{2c_2\{(x-\alpha)^2+\beta^2\}} \frac{\gamma}{c_2\beta} \frac{\mathbf{P}}{\{(x-\alpha)^2+\beta^2\}}.$$

Интегрируя [2], получаем

$$\ln f = -\frac{1}{2c_2}\ln\{(x-\alpha)^2+\mathrm{p}^2\} - \frac{\gamma}{c_2\beta}\operatorname{arctg}\left(\frac{x-\alpha}{\beta}\right) + \text{const},$$

[2] Заметим, что

$$\frac{d}{dx}\ln g(x) = g'(x)/g(x) \quad \text{и} \quad \frac{d}{dx}\operatorname{arctg}\left(\frac{x-\alpha}{\beta}\right) = \beta/\{(x \_ \alpha)^2+\beta^2\}.$$

Таким образом,

$$f = f_0\{(x-\alpha)^2+\beta^2\}^{-1/(2c_2)}\left[\exp\left\{\operatorname{arctg}\left(\frac{x-\alpha}{\beta}\right)\right\}\right]^{-\gamma/(c_2\beta)}.$$

Полученная функция представляет второй из основных типов кривых Пирсона (тип IV). Этот тип наиболее труден для анализа. Диапазон значений распределения неограничен в обе стороны. При переходе к исходной системе координат (от системы, начало которой было смещено в точку среднего значения) частотная функция приобретает вид

$$f_0\{(x-m_1'-\alpha)^2+\beta^2\}^{-1/(2c_2)}\left[\exp\left\{\operatorname{arctg}\left(\frac{x-m_1'-a}{\beta}\right)\right\}\right]^{-\gamma/(c_2\beta)}.$$

Известны значения всех констант, входящих в данную формулу, кроме константы f_0. Найти ее значение непросто. Один из методов предложен в работе [20, с. 59], однако следующий числовой метод представляется более удачным.

Предполагая, что f_0 равно единице, с помощью приближенной формулы интегрирования, например формулы Симпсона (см. параграф 5.3), определим площадь под данной кривой. В рассматриваемом случае эта площадь приблизительно равна 33,475 781 26. Требуется найти такое значение f_0, при котором площадь под кривой равна единице, поэтому мы выбираем $f_0 = 0{,}029\,87$. Подобранная кривая плотности сопоставлена с данными наблюдений в табл. 11.2.3 и на рис. 11.2.2. Площадь участков, лежащих под теми или иными отрезками кривой, может быть найдена методами приближенного интегрирования (см. гл. 5).

Данные в табл. 11.2.3 не являются независимыми наблюдениями значений какой-либо одномерной случайной величины, поэтому критерии согласия, описанные в параграфах 12.11 и 12.12, в данном случае применяться не могут.

Пример **11.2.3.** Подберем кривую Пирсона к данным, приведенным в табл. 11.2.4. Как и ранее, вычисляем

$$m_1' = 34{,}021\,739\,13;$$
$$m_2' = 1\,258{,}695\,652; \quad m_2 = 101{,}216\,9187; \quad \sqrt{b_1} = 0{,}877\,036\,700;$$
$$m_3' = 50\,603{,}260\,86; \quad m_3 = 893{,}0945420; \quad b_1 = 0{,}769\,193\,373;$$
$$m_4' = 2\,209\,891{,}304; \quad m_4 = 45657{,}182\,70; \quad b_2 = 4{,}456\,592\,092;$$
$$a = 3{,}795\,297\,465; \quad d = 3{,}795\,297\,465;$$
$$c_0 = 90{,}609\,144\,67; \quad c_2 = 0{,}034\,934\,126.$$

В данном случае квадратное уравнение, соответствующее знаменателю, имеет два вещественных корня, а именно —35,425 482 16

Таблица 11.2.3. Следующая группа данных из работы У. П. Элдертона. Приведены также соответствующие значения для сглаживания кривой

Центральное значение возрастной группы	Наблюдавшаяся частота	Наблюдавшаяся частота, деленная на 5	Сглаженное значение в центре интервала X 9154
5	10	2,0	0,9
10	13	2,6	3,0
15	41	8,2	9,3
20	115	23,0	25,9
25	326	65,2	62,9
30	675	135,0	129,6
35	1113	222,6	222,1
40	1528	305,6	309,9
45	1692	338,4	346,3
50	1530	306,0	306,7
55	1122	224,4	214,6
60	610	122,0	119,1
65	255	51,0	53,1
70	86	17,2	19,3
75	26	5,2	5,9
80	8	1,6	1,5
85	2	0,4	0,4
90	1	0,2	0,1
95	1	0,2	0,0
	Всего 9154		

Источник. [20, с. 60].

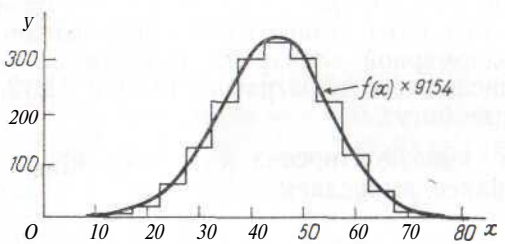


Рис. 11.2.2. Применение пирсоновской кривой типа IV для сглаживания некоторого набора частотных данных из работы У. П. Элдертона

и —73,216 064 07. Отсюда следует, что правую часть дифференциального уравнения можно переписать в виде

$$\frac{A}{35{,}425482\ 16 + x} + \frac{B}{73{,}21606407 + x},$$

Для нахождения значений A и B воспользуемся методом, описанным в примере 11.2.1; подставим значения $x = 0$ и $x = 10$. Получаем систему

$$0{,}028\ 228\ 267A + 0{,}013\ 658\ 204B = -0{,}041\ 886\ 472;$$
$$0{,}022\ 014\ 075A + 0{,}012\ 016\ 9105 = -0{,}104\ 465\ 880,$$

отсюда $A = 23{,}958\ 974\ 99$ и $B = -52{,}584\ 279\ 41$. При применении метода, описанного в примере 11.2.1, для проверки решения подставляем значение $x = 100$ и получаем две близкие оценки значения выражения в правой части: —0,126 659 930 и —0,126 659 928.

Возвращаясь к дифференциальному уравнению, переписываем его в виде

$$\frac{d}{dx}\ln f = \frac{A}{\alpha + x} + \frac{B}{\beta + x},$$

откуда

$$\ln f = A \ln (\alpha + x) + B \ln (\beta + x) + \text{const}$$

$$\left(\text{так как } \frac{d}{dx}\ln (\alpha + \beta x) = \beta/(\alpha + \beta x)\right)$$

$$\text{и } f = f_0 (\alpha + x)^A (\beta + x)^B.$$

Полученная функция представляет третий из основных типов кривых Пирсона (тип VI). Оба корня рассмотренного квадратного уравнения вещественны и имеют одинаковый знак. Распределение задано в диапазоне от $-\alpha$ до ∞ $(\alpha < \beta)$. Значение константы f_0 должно быть выбрано так, чтобы общая площадь под кривой была равна единице. Итак,

$$f_0 = \Gamma(-B)/\{(\beta - \alpha)^{A+B+1}\Gamma(A+1)\Gamma(-B-A-1)\}.$$

Применяя асимптотическую формулу для значений $\Gamma(n)$, приведенную в сноске на с. 132, находим, что $\ln f_0 = 142{,}9889256$.

При переходе к исходной системе координат (от системы, начало которой было смещено в точку среднего значения) частотная функция приобретает вид

$$f_0 (\alpha - m_1' + x)^A (\beta - m_1' + x)^B.$$

Значения переменной x лежат в диапазоне от $m_1' - \alpha$ до $+\infty$. Полученная кривая сопоставлена с исходными данными в табл. 11.2.4 и на рис. 11.2.3. Площадь участков, лежащих под теми или иными

Таблица 11.2.4. Следующая группа данных из работы У. П. Элдертона. Приведены также соответствующие значения для сглаживающей кривой

Центральное значение группы десяти возрастов	Наблюдавшаяся частота	Наблюдавшаяся частота, деленная на 10	Сглаженное значение в центре интервала X 368
10	1	0,1	0,03
20	56	5,6	6,06
30	167	16,7	16,09
40	98	9,8	10,01
50	34	3,4	3,44
60	9	0,9	0,91
70	2	0,2	0,22
80	1	0,1	0,05
	Всего 368	—	—

Источник. [20, с. 68].

отрезками кривой, может быть найдена методами приближенного интегрирования (см. гл. 5).

Данные, приведенные в табл. 11.2.4, не являются независимыми наблюдениями значений какой-либо одномерной случайной вели-

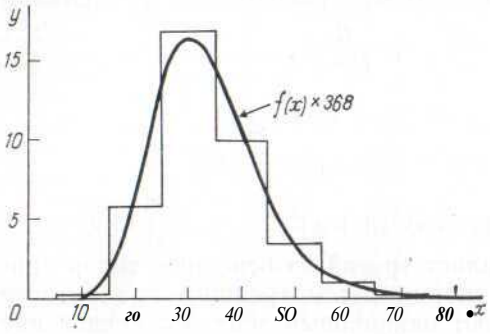


Рис. 11.2.3. Применение пирсоновской кривой типа VI для сглаживания некоторого набора частотных данных из работы У. П. Элдертона

чины, поэтому критерии согласия, описанные в параграфах 12.11 и 12.12, в данном случае применяться не могут.

Литература: [11, с. 387], [20, с. 47—109, 182—184], [21, с. 50-52; русский перевод с. 64—67].

11.3. ПОПРАВКИ НА СГРУППИРОВАННОСТЬ ДАННЫХ

В каждом из трех примеров, приведенных в параграфе 11.2, рассматривались сгруппированные данные. Некоторые авторы полагают, что при вычислении значений моментов необходимо ввести поправки, которые учитывали бы группировку. В случае когда обеим хвостовым частям распределения соответствует функция плотности, значения которой плавно стремятся к нулю при неограниченном росте или неограниченном убывании значений случайной величины, иногда используют поправки Шеппарда (h обозначает величину интервала значений для каждой из групп данных):

$$m_2\,(\text{исправленное}) = m_2 - h^2/12, \qquad (11.3.1)$$

$$m_4\,(\text{исправленное}) = m_4 - h^2 m_2/2 + 7h^4/240. \qquad (11.3.2)$$

Для моментов m_1' и m_3 поправки не нужны.

Литература: [20, с. 177—181], [53, с. 75—81; русский перевод с. 111—117], [93, с. 90].

11.4. УПРАЖНЕНИЯ

1. В примере 11.2.2 значение характеристики асимметрии $\sqrt{b_1}$ близко к нулю, а значение характеристики эксцесса b_2 близко к трем. К приведенным в таблице данным подберите кривую нормального распределения, приравняв значения μ и σ^2 значениям выборочных моментов m_1' и m_2. Сравните полученный результат с результатом подбора кривой Пирсона (тип IV).

2. С помощью формулы Стирлинга, приведенной в сноске на с. 132, получите приближенные значения $n!$ ($n=1, 2, \ldots, 5$). Подсчитайте для каждого из этих случаев относительную погрешность приближения.

12. ПРОВЕРКА ГИПОТЕЗ

В данной главе описана общая процедура проверки статистических гипотез, а также рассмотрено большое число стандартных критериев. Эти критерии включают проверку гипотез о биномиальном и полиномиальном распределении (параграфы 12.6—12.8), зависимости в таблицах сопряженности признаков (параграфы 12.9 и 12.10), согласия распределений (параграфы 12.11 и 12.12), гипотез о выборочных средних (параграфы 12.13—12.27), гипотез о величине дисперсий (параграфы 12.28—12.30), гипотез о корреляции (параграфы 12.31—12.34) и форме распределения (параграфы 12.35—12.37).

12.1. ВВЕДЕНИЕ

Предположим, что исследователь получил монету и хочет проверить, симметрична ли она. Очевидно, что он должен провести следующий эксперимент: подбросить монету много раз и рассмотреть число выпадений герба и решетки. Предположим, что он решил подбросить ее 1000 раз. Для симметричной модели $p = \frac{1}{2}$ и ожидаемое число выпадений герба равно при таком эксперименте 500 (см. параграф 10.1). Если в результате окажется, что герб выпал 511 раз, то исследователь, по-видимому, сделает вывод о симметричности монеты (так как 511 близко к 500); если же окажется, что герб выпал 897 раз, то он будет совершенно убежден в том, что монета несимметрична, и отбросит предположение о ее симметричности.

Вероятность выпадения герба точно 511 раз мала, вероятность выпадения герба точно 897 раз тоже мала. Почему же тогда мы принимаем нулевую гипотезу о симметричности монеты в первом случае (511 выпадений) и отклоняем ее во втором (897 выпадений)? При выпадении герба 511 раз мы принимаем *нулевую гипотезу* о симметричности, поскольку вероятность того, что при подбрасывании симметричной монеты число выпадений герба будет отличаться от среднего значения 500 на 11 или более единиц, довольно велика (около 0,5). Напротив, вероятность того, что в аналогичном эксперименте число выпадений герба будет отличаться от среднего значения 500 на 397 или более единиц, ничтожно мала, поэтому в данном случае нулевая гипотеза отклоняется.

Подобные рассуждения лежат в основе всех статистических проверок. Нулевая гипотеза отклоняется, если вероятность того, что она верна, оказывается ниже некоторого уровня, называемого

уровнем значимости. Дадим краткое описание типичных этапов процедуры проверки статистических гипотез.

Типичная процедура	*Пример с подбрасыванием монеты*
1. Выбрать *уровень значимости* α.	$\alpha = 0{,}05$.
2. Описать *статистическую модель*.	Число выпадений герба X подчиняется биномиальному распределению с параметрами $n = 1000$ и p.
3. Сформулировать *нулевую гипотезу* H_0 и *альтернативную гипотезу* H_1.	$H_0 : p = \frac{1}{2}$, $H_1 : p \neq \frac{1}{2}$.
4. Выбрать критериальную статистику (критерий)*, поведение которой известно.	$T = (X - np)/\sqrt{npq}$ при росте n является случайной величиной, распределение которой приближается к стандартному нормальному распределению.
5. Для проверки нулевой гипотезы определить подходящую *критическую область*. (Если значение принятого статистического критерия попадает в эту область, то нулевая гипотеза отклоняется). Вероятность попадания значения этого критерия в указанную область при условии, что H_0 справедлива, равна α.	По таблицам нормального распределения устанавливается, что к критической области относятся такие значения, которые удовлетворяют неравенству $\lvert T \rvert > 1{,}96$.
6. Вычислить значение статистического критерия.	$T = (511 - 500)/\sqrt{250} = 0{,}696$.
7. Сделать выводы. Если полученное значение критерия лежит в критической области, то следует отклонить нулевую гипотезу и принять альтернативную. В противном случае принять нулевую гипотезу.	Данное значение критерия не попадает в критическую область, поэтому нулевая гипотеза принимается. Нет никакого основания считать монету несимметричной.

Тот факт, что полученное значение критерия незначимо, не является доказательством справедливости нулевой гипотезы; он лишь показывает, что имеющиеся данные ей не противоречат. Нулевая гипотеза принимается до тех пор, пока не будет получено доказательство ее ложности. В более общем контексте подобный подход называют *научным методом*.

Литература: [2, с. 27—32; русский перевод с. 42—49], [3, с. 12], [7, с. 97—104; русский перевод с. 77—84], [16, с. 86—96], [39, с. 1—14], [45, с. 159—164], [50, с. 195—202; русский перевод с. 215—220], [70, с. 275—290], [81, с. 76—83], [86. с. 193—199], [91, с. 24—25; русский перевод с. 42—44], [93, с. 167—187], [102, с. 216—219].

12.2. ТИПЫ ОШИБОК И МОЩНОСТЬ КРИТЕРИЯ

Как было объяснено, *критическая область* — это множество таких исходов некоторого статистического эксперимента, которые

* Иногда употребляют термин *тестовая статистика* или *статистика критерия*.— *Примеч. пер.*

приводят нас к отклонению нулевой гипотезы. Если нулевая гипотеза справедлива и принят уровень значимости α, то с вероятностью α исход эксперимента попадет в критическую область; в этом случае нулевая гипотеза будет ошибочно отклонена и тем самым будет допущена *ошибка I рода*. Иногда возникает ситуация, когда нулевая гипотеза ложна, а исход эксперимента не попал в критическую область. В таком случае нулевая гипотеза ошибочно принимается и тем самым допускается *ошибка II рода*.

Естественно стремление минимизировать вероятности ошибок рассмотренных типов. Снижая уровень значимости α, можно легко сократить вероятность возникновения ошибки I рода, но в таком случае возрастет вероятность ошибки II рода. В связи с этим вводят понятие *мощности* критерия, которую определяют как вероятность отклонений нулевой гипотезы. Эта вероятность зависит от реального значения рассматриваемого параметра совокупности. Поскольку это значение заранее неизвестно, рассматривают *кривую мощности*, которая показывает соответствующее значение мощности критерия для каждого возможного значения параметра. Очевидно, что точки идеальной кривой мощности имеют ординату, равную единице, для всех значений параметра, кроме тех, которые соответствуют нулевой гипотезе. На практике получить подобный результат невозможно, но обычно удается повысить мощность критерия до любого желаемого уровня, соответственно увеличив объем выборки.

Следует предупредить об опасности, связанной с применением нескольких статистических критериев при анализе одних и тех же данных. Если к одним и тем же данным применяют два различных критерия для проверки одной и той же нулевой гипотезы (или двух сходных гипотез) и в каждом случае принимается уровень значимости, равный, например, 5 %, то вероятность того, что хотя бы по одному из критериев нулевая гипотеза будет ошибочно отклонена, превосходит 5 %. Следует воспользоваться лишь одним критерием, желательно более мощным.

Иногда возникает необходимость проверки двух различающихся гипотез при использовании одних и тех же данных (например, гипотезы о значениях среднего и дисперсии некоторой нормально распределенной совокупности). Если для обоих критериев принимается 5 %-ный уровень значимости и обе нулевые гипотезы справедливы, то вероятность ошибочного отклонения хотя бы одной из нулевых гипотез значительно превосходит 5 % и часто бывает близка к 10 %.

Пример 12.2.1. В примере с подбрасыванием монеты, описанном в параграфе 12.1, нулевая гипотеза отклонялась во всех случаях, когда $\lvert T \rvert > 1{,}96$. Это значит, что мы отклоняем нулевую гипотезу во всех случаях, когда число выпадений герба лежит вне диапазона от 470 до 530 (включительно).

Предположим, что монета на самом деле немного несимметрична и что неизвестное нам значение p в действительности равно

0,52. Тогда в соответствии с (10.2.2) вероятность отклонения нулевой гипотезы $p = 0{,}5$ равна:

$$1 - \Phi\left(\frac{530{,}5 - 1000 \times 0{,}52}{\sqrt{1000 \times 0{,}52 \times 0{,}48}}\right) + \Phi\left(\frac{469{,}5 - 1000 \times 0{,}52}{\sqrt{1000 \times 0{,}52 \times 0{,}48}}\right) = 0{,}254.$$

Таково значение мощности критерия при $p = 0{,}52$; вероятность совершить ошибку II рода равна 0,746. Полностью кривая мощности показана на рис. 12.2.1. Там же показана кривая мощности данного критерия для выборки объемом 10 000.

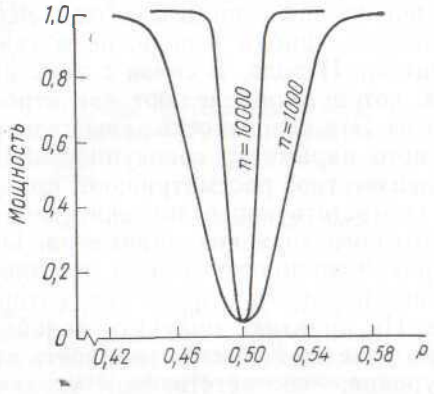


Рис. 12.2.1. Кривые мощности, соответствующие значениям $n = 1000$ и $n = 10\,000$ для критерия, проверяющего биномиальную гипотезу $p = 0{,}5$. В обоих случаях уровень значимости равен 5 %

Литература: [50, с. 195—202; русский перевод с. 215—220], [93, с. 167—187].

12.3. ОДНОСТОРОННИЕ И ДВУСТОРОННИЕ КРИТЕРИИ

Предположим, что игрок достал монету, которую он собирается использовать в игре с ничего не подозревающим партнером. Ему хотелось бы проверить заверения изготовителя, что вероятность выпадения герба для данной монеты немного меньше половины. Он подбрасывает монету 1000 раз, получая при этом 529 выпадений герба. Противоречит ли этот результат заверениям изготовителя?

Утверждение изготовителя будет признано ложным лишь в том случае, когда в результате эксперимента обнаружится, что число выпадений герба слишком велико (односторонний критерий). В предыдущем примере нулевая гипотеза о симметричности монеты отклоняется в том случае, когда число выпадений герба либо слишком мало, либо слишком велико (двусторонний критерий). В параграфе 12.6 описаны оба эти критерия.

12.4. УСТОЙЧИВОСТЬ. НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ

Все статистические критерии предполагают некоторый конкретный тип математической модели (например, нормально распределенная совокупность с заданным значением дисперсии).

На практике условия, налагаемые математической моделью, могут и не выполняться, что приводит к возрастанию вероятности неправильных выводов, которые делают на основе рассматриваемого критерия. Для одних критериев подобное снижение надежности выводов происходит в большей степени, чем для других. *Устойчивыми (робастными)* называют такие критерии, для которых умеренные отклонения совокупности от предполагаемой математической модели мало влияют на надежность выводов *.

Многие из критериев, описанные в данной главе, основаны на предположении о нормальной распределенности совокупности. Критерии гипотез о средних значениях в большинстве своем устойчивы при умеренных отклонениях от нормальности, а критерии для гипотез о значениях дисперсии — нет.

Некоторые из критериев, описанных в данной главе, являются *непараметрическими* (или *свободными от распределения*). Применение таких критериев не основывается на предположениях о каком-либо конкретном виде распределения, соответствующего изучаемой совокупности. При рассмотрении нормально распределенной совокупности критерии этого типа несколько уступают по мощности соответствующим критериям, построенным на предположении о нормальности; они обладают, однако, тем преимуществом, что свободны от подобного предположения о нормальности, поэтому их можно использовать в ситуациях, когда вид распределения заранее неизвестен. Это соответствует общей закономерности: статистическая модель, связанная с введением дополнительных предположений о распределении, как правило, позволяет построить более мощный критерий.

Литература: [3, с. 219—221], [16, с. 96—99], [50, с. 255—258; русский перевод с. 314—319], [54, с. 470—473; русский перевод с. 622—633].

12.5. СПОСОБ ОПИСАНИЯ КРИТЕРИЕВ, ПРИНЯТЫЙ В ДАННОЙ ГЛАВЕ

Описание каждого критерия разбито на шесть пунктов.

Обычная форма записи данных. Этот пункт позволяет читателю представить себе общий вид записи данных, к анализу которых *мог бы быть* применен рассматриваемый критерий.

Статистическая модель. *Трудно переоценить значение этого пункта.* Если модель, на которой основан критерий, не соответствует рассматриваемым данным, то весьма вероятно, что те выводы, которые можно сделать, применяя описываемый критерий, окажутся ошибочными. Предположения о независимости испытаний обязательно должны отвечать действительности. С другой стороны, часто можно пренебречь умеренными отклонениями от нормальности (см. параграф 12.4).

---

* О проблеме устойчивости см.: Смоляк С. А., Титаренко Б. П. Устойчивые методы оценивания. М., Статистика, 1980.— *Примеч. пер.*

Гипотезы. Приводятся обычные нулевая и альтернативная гипотезы.

Критическая область. При описании всех критериев предполагается, что выбран 5 %-ный уровень значимости. Не представляет никаких трудностей переход к любому другому уровню значимости.

Вычисление значения критериальной статистики. Описание вычислительной процедуры.

Комментарии. Различные комментарии по вопросам вычислений, альтернативных критериев, других нулевых гипотез, устойчивости и т. д.

12.6. КРИТЕРИЙ, В КОТОРЫЙ ВХОДИТ ЕДИНСТВЕННАЯ БИНОМИАЛЬНАЯ ВЕРОЯТНОСТЬ p

Обычная форма записи данных.

Число испытаний, закончившихся		Общее число испытаний
успехом	неудачей	
r	$n-r$	n

Статистическая модель. n испытаний независимы и для каждого из них вероятность успеха равна p. Следовательно, распределение числа успешных испытаний является биномиальным.

Гипотезы.

а) *Равенство*

$H_0: p = p_0;$

$H_1: p \neq p_0.$

б) *Неравенство*

$H_0: p \leqslant p_0;$

$H_1: p > p_0.$

в) *Неравенство*

$H_0: p \geqslant p_0;$

$H_1: p < p_0.$

Критическая область.

а) *Равенство.* Значения, лежащие выше верхней 2,5 %-ной или ниже нижней 2,5 %-ной точек стандартного нормального распределения.

б) *Неравенство.* Значения, лежащие выше верхней 5 %-ной точки стандартного нормального распределения.

в) *Неравенство.* Значения, лежащие ниже нижней 5 %-ной точки стандартного нормального распределения.

Вычисление значения критериальной статистики. Подставить $p = p_0$ в формулу

$$T = (r - np)/\{np(1-p)\}^{\frac{1}{2}}. \qquad (12.6.1)$$

Комментарии. 1. Описанный критерий основан на аппроксимации биномиального распределения нормальным (см. параграф 10.2), поэтому его можно применять лишь при больших значениях n (например, $n > 20$). Для меньших значений n в качестве статистического критерия можно использовать величину r и соответственно следующие критические области:

а) *Равенство.* Верхняя и нижняя 2,5 %-ные области значений биномиального распределения (n, p_0).

б) *Неравенство.* Верхняя 5 %-ная область значений биномиального распределения (n, p_0).

в) *Неравенство.* Нижняя 5 %-ная область значений биномиального распределения (n, p_0).

2. Неудобства, связанного с использованием таблиц биномиального распределения, можно избежать, прибегнув к следующим процедурам проверки (основанным на формулах (10.3.2) и (10.3.3)), применимым при любых значениях:

а) *Равенство.* Вычислить

$$Y = \frac{r}{n+1-r}\,\frac{1-p_0}{p_0}, \qquad (12.6.2)$$

$$Z = \frac{n-r}{r+1}\,\frac{p_0}{1-p_0}. \qquad (12.6.3)$$

Нулевая гипотеза отклоняется, если полученное значение величины Y превосходит верхнее 2,5 %-ное значение F-распределения с $2(n+1-r)$ и $2r$ степенями свободы или если значение величины Z превосходит верхнее 2,5 %-ное значение F-распределения с $2(r+1)$ и $2(n-r)$ степенями свободы.

б) *Неравенство.* Применить формулу (12.6.2) для вычисления значения Y. Если это значение лежит выше 5 %-ной точки F-распределения с $2(n+1-r)$ и $2r$ степенями свободы, то нулевая гипотеза отклоняется.

в) *Неравенство.* Применить (12.6.3) для вычисления значения Z. Нулевая гипотеза отклоняется, если это значение лежит выше 5 %-ной точки F-распределения с $2(r+1)$ и $2(n-r)$ степенями свободы.

Пример 12.6.1. Известно, что рассматриваемая совокупность распределена по нормальному закону с единичной дисперсией. Среднее значение неизвестно. Из этой совокупности взята выборка объемом 35: двадцать наблюдений превышают 10 и пятнадцать наблюдений лежат ниже 10. Разумно ли утверждать, что среднее значение равно 9?

Если бы среднее значение совокупности было равно 9, то для каждого из наблюдений вероятность оказаться меньше 10 была бы равна: $p = \Phi(10-9)/\sqrt{1}) = 0{,}8413$, а вероятность оказаться больше 10 — 0,1587. Следовательно, необходимо проверить нулевую гипотезу H_0: $p = 0{,}8413$ при альтернативной гипотезе $H_1: p \neq 0{,}8413$. Объем выборки достаточно велик, поэтому можно воспользоваться статистическим критерием (12.6.1) и рассмотреть критическую область, соответствующую обоим «хвостам» нормаль-

ной кривой. 5 %-ная критическая область задается неравенством $|T|>1{,}96$. Значение критерия в данном примере равно:

$$T = (15 - 35 \times 0{,}8413)/(35 \times 0{,}8413 \times 0{,}1587)^{\frac{1}{2}} = -6{,}68.$$

Это значение лежит в критической области, поэтому мы отклоняем нулевую гипотезу и делаем вывод, что $p \neq 0{,}8413$.

Можно также применить метод, описанный в п. 2 комментариев. Вычисляем

$$Y = \frac{15}{35+1-15}\,\frac{0{,}1587}{0{,}8413} = 0{,}13;$$

$$Z = \frac{35-15}{15+1}\,\frac{0{,}8413}{0{,}1587} = 6{,}63.$$

Z существенно превосходит 2,5 %-ное значение $F_{32,40}$-распределения, следовательно, мы отклоняем нулевую гипотезу и делаем вывод, что $p \neq 0{,}8413$.

Литература: [3, с. 113], [7, с. 146—150; русский перевод с. 135—140], [45, с. 175—178], [105, с. 287—290].

12.7. ПРОВЕРКА ГИПОТЕЗЫ ОТНОСИТЕЛЬНО ЕДИНСТВЕННОГО РЯДА ПОЛИНОМИАЛЬНЫХ ВЕРОЯТНОСТЕЙ

Обычная форма записи данных.

Число испытаний, завершившихся исходом типа	Общее число испытаний
1 2... k	
n_1 n_2... n_k	$n = n_1 + \ldots + n_k$

Статистическая модель. Проводится n независимых испытаний, для каждого из них вероятность исхода i-го типа равна p_i $(i = 1, 2, \ldots, k)$. Таким образом, распределение предполагается полиномиальным.

Гипотезы. H_0: p1 $= \pi_1$, $p_2 = \pi_2$, ..., $p_k = \pi_k$; H_1: нулевая гипотеза неверна.

Критическая область. Значения выше 5 %-ной точки распределения χ^2 с $k-1$ степенями свободы.

Вычисление значения критериальной статистики. Подставим значения $p\backslash - \pi_1, \ldots, p_k = \pi_k$ в формулу критерия

$$\chi^2 = (n_1 - np_1)^2/(np_1) + \ldots + (n_k - np_k)^2/(np_k). \qquad (12.7.1)$$

Комментарии. 1. Критерий основан на аппроксимации полиномиального распределения распределением χ^2 (см. параграф 10.5), поэтому он может применяться лишь при больших значениях n (например, $n>25$). Математические ожидания $\{n\pi_i\}$ исходов для каждого типа не должны быть слишком малы. Многие авторы предлагают в качестве минимального допустимого значения математического ожидания для какого-либо типа исходов число 5, хотя Кокрен [13] считает, что это минимальное значение может быть равно единице при условии, что лишь для небольшого числа типов математическое ожидание меньше 5 (одна клетка из пяти или две из десяти и более клеток).

2. Если математическое ожидание, соответствующее какой-либо конкретной клетке, слишком мало, то эту клетку следует объединить с какой-нибудь другой клеткой, уменьшив тем самым число степеней свободы на единицу. (При этом необходимо сложить соответствующие математические ожидания и аналогичным образом сложить числа наблюдений.) Если нет необходимости группировать типы исходов, то группировку производить нецелесообразно, так как это снижает чувствительность критерия.

3. Этот критерий может применяться вместо критерия а), описанного в параграфе 12.6. Для случая биномиального распределения эти два критерия эквивалентны.

4. Этот критерий можно применять для проверки гипотезы о том, что рассматриваемая выборка $x_1, x_2, \ldots, x_n$ объема n взята из некоторой конкретной совокупности, которой соответствует функция распределения $F(x)$. Диапазон возможных значений данной случайной величины разбивают на k непересекающихся интервалов $(-\infty, a_1), (a_1, a_2), \ldots, (a_{k-1}, \infty)$. В рамках нулевой гипотезы вероятность попадания отдельного наблюдения в i-й интервал равна: $pi - F(a_i) - F(a_{i-1})$ (см. параграф 8.2). Производится обычная процедура проверки по критерию χ^2 (см. пример 12.7.3). Другой метод проверки описан в параграфе 12.12.

Пример 12.7.1. Шестигранная кость была брошена 50 раз, результаты этого испытания таковы: 12 раз выпала «шестерка», 9 — «пятерка», 9 — «четверка», 6 — «тройка», 9 — «двойка» и 5 — «единица». Можно ли на основании этих результатов утверждать, что кость несимметрична?

Необходимо проверить нулевую гипотезу

$$H_0: p_1 = p_2 - P_3 = p_4 = p_5 = p_6 = \frac{1}{6};$$

H_1: нулевая гипотеза неверна.

Критической областью в данном случае является верхняя Г) %-ная область распределения χ^2_5; она включает значения, превышающие 11,070. Объем выборки достаточно велик, поэтому можно применить статистический критерий (12.7.1):

$$\chi^2_5 = \left(12 - 50 \times \frac{1}{6}\right)^2 \Big/ \left(50 \times \frac{1}{6}\right) + \ldots$$

$$\ldots + \left(5 - 50 \times \frac{1}{6}\right)^2 \Big/ \left(50 \times \frac{1}{6}\right) = 3{,}76.$$

Полученный результат незначим и нет никакого подтверждения несимметричности кости.

Пример 12.7.2[1]. В своей статье по генетике Батесон и Паннетт приводят данные о цвете и форме 427 частиц пыльцы (см. [78]). Эти данные воспроизведены в табл. 12.9.1. Проверим нулевую гипотезу о том, что гены, обусловливающие пурпурный цвет частиц и их вытянутую форму, являются доминантными и что отсутствует связь между распределениями по цвету и форме.

В рамках данной нулевой гипотезы каждая частиица с вероятностью $\frac{3}{4}$ имеет пурпурную окраску и с вероятностью $\frac{1}{4}$ -- красную, с вероятностью $\frac{3}{4}$ частица имеет вытянутую форму и с вероятностью $\frac{1}{4}$ — круглую. Отсутствие связи предполагает независимость указанных признаков, поэтому проверяемая нулевая гипотеза формулируется следующим образом:

$$H_0: \begin{cases} P\ (\text{частица пурпурная и вытянутая}) = \frac{9}{16}, \\ P\ (\text{частица пурпурная и круглая}) = \frac{3}{16}, \\ P\ (\text{частица красная и вытянутая}) = \frac{3}{16}, \\ P\ (\text{частица красная и круглая}) = \frac{1}{16}. \end{cases}$$

5 %-ное значение распределения χ^2_3 равно 7,81. В соответствии с нулевой гипотезой ожидаемое число пурпурных/вытянутых частиц равно: $\frac{9}{16} \times 427 = 240{,}19$; ожидаемое число пурпурных/круглых частиц равно: $\frac{3}{16} \times 427 = 80{,}06$; ожидаемое число красных/вытянутых частиц равно: $\frac{3}{16} \times 427 = 80{,}06$; ожидаемое число красных/круглых частиц равно: $\frac{1}{16} \times 427 = 26{,}69$ Значение критерия таково:

$$\chi^2 = (296 - 240{,}19)^2/240{,}19 + (19 - 80{,}06)^2/80{,}06 + (27 - 80{,}06)^2/80{,}06 + (85 - 26{,}69)^2/26{,}69 = 222{,}1.$$

[1] Сравните этот пример с примером 12.9.1.

Это значение (высоко) значимо, поэтому есть все основания отклонить нулевую гипотезу.

Пример 12.7.3. Следующие двадцать наблюдений взяты из неизвестной совокупности:

0,33; -0,52; -2,41; -1,93; 0,46; -0,44; -0,97;
-0,38; 0,48; 1,29; -1,82; -1,23; -0,21; 2,66.
—1,22; -0,41; -0,95; 1,47; —0,83; -0,43;

Проверим нулевую гипотезу о том, что рассматриваемое распределение нормально с нулевым средним и единичной дисперсией.

Для того чтобы воспользоваться критерием χ^2, удобно разделить диапазон возможных значений нормального распределения па шесть интервалов: $(-\infty; -1{,}5)$, $(-1{,}5; -0{,}5)$, $(-0{,}5; 0)$; $(0; 0{,}5)$, $(0{,}5; 1{,}5)$ и $(1{,}5; \infty)$. Ожидаемое число наблюдений, попадающих, например, в интервал (0,5; 1,5) равно: $20(\Phi(1{,}5) - \Phi(0{,}5)) = 4{,}8346$. Аналогично получены ожидаемые значения числа попаданий и для всех остальных интервалов (см. табл. 12.7.1).

Таблица 12.7.1. **Проверка нулевой гипотезы о нормальности исследуемого распределения (с нулевым средним и единичной дисперсией) методом χ^2**

Интервал	Число наблюдений в данном интервале		Вклад в значение χ^2
	„фактическое"	ожидаемое	
От $-\infty$ до —1,5	3	1,3362	2,0717
От -1,5 до -0,5	6	4,8346	0,2809
От —0,5 до 0,0	5	3,8292	0,3580
От 0,0 до 0,5	3	3,8292	0,1796
От 0,5 до 1,5	2	4,8346	1,6620
От 1,5 до со	1	1,3362	0,0846
Всего	20	20,0000	4,6368

Г) %-ное критическое значение для распределения χ^2 с пятью степенями свободы равно 11,07. Полученное значение χ^2 (4,64) существенно ниже этого уровня, поэтому нулевая гипотеза принимается и можно сделать вывод о нормальности исследуемого распределения (с нулевым средним и единичной дисперсией).

Читателю необходимо обратить внимание на пример 12.12.1.

Литература: [7, с. 140—144; русский перевод с. 130—133], [9, С. 305], [13], [86, с. 232—233].

12.8. РАВЕНСТВО ПОЛИНОМИАЛЬНЫХ (БИНОМИАЛЬНЫХ) ВЕРОЯТНОСТЕЙ В ДВУХ ИЛИ БОЛЕЕ ЭКСПЕРИМЕНТАХ. РАЗНОСТЬ МЕЖДУ ДВУМЯ БИНОМИАЛЬНЫМИ ВЕРОЯТНОСТЯМИ

Обычная форма записи данных.

	Число испытаний, завершившихся исходом типа				Общее число испытаний
	1	2	...	k	
Эксперимент 1	n_{11}	«12	...	n_{1k}	$n_{1.}$
Эксперимент 2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
Эксперимент r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r.}$
Все эксперименты	$n_{.1}$	И.2	...	$n_{.k}$	$n_{..}$

Примечание. $n_{1.} = n_{11} + n_{12} + \ldots + n_{1k}$,

$n_{.1} = n_{11} + n_{21} + \ldots + n_{r1}$,

$n_{..} = \text{л.1} + \bullet\bullet\bullet + n_{.k} = \text{л1.} + \ldots + n_{r.}\bullet$

Статистическая модель. Рассматриваемые $n_{..}$ испытаний предполагаются независимыми. В эксперименте 1 вероятность исхода 1 равна p_{11}, вероятность исхода 2 равна $p_{12}, \ldots$; в эксперименте 2 вероятность исхода 1 равна p_{21}, вероятность исхода 2 равна $p_{22} \ldots$; и т. д. Таким образом, для каждого эксперимента распределение исходов является полиномиальным.

Гипотезы.

$$H_0: \begin{cases} p_{11} = p_{21} = \ldots = p_{r1}, \\ p_{12} = p_{22} = \ldots = p_{r2}, \\ \ldots\ldots\ldots\ldots \\ p_{1k} = p_{2k} = \ldots = p_{rk}; \end{cases}$$

H_1: нулевая гипотеза неверна.

Критическая область. Значения, превышающие верхнюю 5 %-ную точку распределения χ^2 с $(r-1)(k-1)$ степенями свободы.

Вычисление значения критериальной статистики. Если нулевая гипотеза верна, то наилучшая оценка $\hat{p}_1$ общего значения величин p_{11}, p_{21}, ..., p_{r1} задается выражением $n_{.1}/n_{..}$. Эксперимент 1 состоит из fti. испытаний, поэтому можно ожидать, что $n_{1.}\hat{p}_1 = n_{1.} \times n_{.1}/n_{..}$ из этих испытаний приведут к исходу типа 1. Эта процедура повторяется для каждой из клеток таблицы, приведенной выше, т. е. перемножаются суммы элементов соответствующих столбцов и строк. Это произведение делится на общее число испытаний. Полученный результат показывает ожидаемое

число испытаний для рассматриваемой клетки. Далее вычисляют значение

$$\chi^2 = \sum_{\substack{\text{по всем} \\ rk \text{ клеткам}}} \frac{(\text{фактическое число} - \text{ожидаемое число})^2}{\text{ожидаемое число}}. \quad (12.8.1)$$

Комментарии. 1. В случае биномиального распределения $k = 2$.

2. Критерий основан на аппроксимации полиномиального распределения распределением χ^2 (см. параграф 10.5), поэтому его можно применять лишь к выборкам достаточно большого объема. Ожидаемые числа для различных клеток не должны быть слишком малыми (см. пункты 1 и 2 комментариев в параграфе 12.7). Если клетка слишком мала и ее нужно объединить с другой, то при этом следует объединять клетки, относящиеся к одному и тому же эксперименту.

3. Эта статистическая модель во многом сходна (хотя и не полностью идентична) с моделью таблиц сопряженности, приведенной в параграфе 12.9.

4. В особом случае двух биномиальных экспериментов рассматриваемое распределение χ^2 имеет одну степень свободы; статистический критерий (т. е. соответствующая функция) является и данном случае квадратом единичной нормальной случайной величины. Для проверки гипотезы H_0, где

$$H_0: p_{11} - p_{21} = \delta \text{ (заданная константа)},$$
$$H_1: p_{11} - p_{21} \neq \delta,$$

следует рассмотреть в качестве критической области объединение верхней и нижней 2,5 %-ных областей стандартного нормального распределения и воспользоваться статистическим критерием

$$T = (\hat{p}_{11} - p_{21} - \delta)/(\hat{p}_{11}\hat{q}_{11}/n_{1.} + \hat{p}_{21}\hat{q}_{21}/n_{2.})^{\frac{1}{2}}, \quad (12.8.2)$$

где

$$p_{11} = n_{11}/n_{1.},\ \hat{q}_{11} = 1 - \text{P11},$$
$$p_{21} = n_{21}/n_{2.},\ \hat{q}_{21} = 1 - \hat{p}_{21}.$$

Для проверки гипотезы H_0, где

$$H_0: p_{11} - p_{21} \leqslant \delta,$$
$$H_1: p_{11} - p_{21} > \delta,$$

следует рассмотреть верхнюю 5 %-ную область стандартного нормального распределения и воспользоваться критерием (12.8.2),

Пример 12.8.1. Проводится исследование о продолжительности пребывания пациентов в больнице после операции определенного типа. От каждого из трех штатов Австралии случайным образом выбрано по пятьдесят пациентов (из Нового Южного Уэльса, Квинсленда и Тасмании).

Продолжительность пребывания в больнице каждого пациента фиксируется с точностью до недели: менее недели, от одной до двух

недель, от двух до трех недель, от трех до четырех недель и свыше четырех недель. Результаты обследования приведены в табл. 12.8.1. Одинакова ли продолжительность пребывания пациентов в больнице для этих трех штатов?

Таблица 12.8.1. Продолжительность пребывания пациентов в больнице

Штат	Продолжительность (недели)					Всего
	<1	1-2	2-3	3-4	>4	
	Число пациентов					
Новый Южный Уэльс	8	16	16	6	4	50
Квинсленд	6	13	14	7	10	50
Тасмания	5	12	10	10	13	50
Всего	19	41	40	23	27	150

Необходимо проверить гипотезу Яо.

H_0: распределения по срокам одинаковы во всех штатах;

H_1: нулевая гипотеза неверна.

Если нулевая гипотеза верна, то ожидаемое число пациентов из Нового Южного Уэльса с продолжительностью пребывания менее одной недели равно: $19 \times 50/150 = 6{,}3$, а ожидаемое число пациентов Тасмании с длительностью пребывания свыше четырех недель равно: $27 \times 50/150 = 9{,}0$. Аналогичные результаты могут быть получены и для тринадцати остальных клеток. 5 %-ной критической областью является верхняя 5 %-ная область распределения χ^2_8; в нее входят значения, превышающие 15,51. Значение статистического критерия в данном случае таково:

$$\chi^2 = (8 - 6{,}3)^2/6{,}3 + \ldots + (13 - 9{,}0)^2/9{,}0 = 8{,}57.$$

Это значение не является значимым, поэтому нет никаких оснований отклонить нулевую гипотезу.

Следует отметить, что данные в таблице упорядочены (пациенты разбиты на группы по продолжительности пребывания в больнице). Примененный критерий не учитывает это обстоятельство, т. е. он не полностью использует имеющуюся информацию.

Литература: [3, с. 115], [9, с. 306—308, 311—312], [13], [45, с. 178—180], [105, с. 296—297].

12.9. ПРОВЕРКА ЗАВИСИМОСТИ В ТАБЛИЦЕ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

Обычная форма записи данных. Нередко случается, что частотные данные одновременно собираются для характеристики двух переменных. Примером может служить информация о цвете глаз

и волос индивидуума. Так, в выборке объема $n_{..}$ n_{ij} людей имеют волосы «класса» i и глаза «класса» j. Такого рода данные могут быть представлены в виде сопряженности признаков:

Класс „цвет волос"	Класс „цвет глаз"				Всего
	1	2	...	s	
1	«11	«12	...	n_{1s}	$n_{1.}$
2	«21	n_{22}	...	n_{2s}	$n_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r.}$
Всего	«Л	$n_{.2}$	...	$n_{.s}$	п..

Статистическая модель. Наблюдаются случайно отобранные люди, общее число которых равно $n_{..}$, с различным цветом волос и глаз. Число людей с цветом волос 2 и цветом глаз 1 обозначается n_{21}. Аналогично обозначение и для других клеток. Таким образом, индивидуум попадает в одну из rs клеток, соответственно распределение по rs клеткам является полиномиальным.

Гипотезы.

Яо: между цветом волос и цветом глаз индивидуума нет зависимости;

H_1: нулевая гипотеза неверна.

Критическая область. Область, лежащая выше верхней 5 %-ной точки распределения χ^2 с $(r-1)(s-1)$ степенями свободы.

Вычисление значения критериальной статистики. Вычислим ожидаемые значения в каждой клетке, умножая n_i-элемент столбца «всего», соответствующий данной клетке, на n_j-элемент строки «всего», соответствующий данной клетке, и деля получившееся произведение на общее число $n_{..}$, как в параграфе 12.8. Далее следует вычислить

$$\chi^2 = \sum_{\substack{\text{по всем } rs \\ \text{клеткам}}} \frac{(\text{наблюдаемое число} - \text{ожидаемое число})^2}{\text{ожидаемое число}}. \qquad (12.9.1)$$

Комментарии. 1. Этот критерий является приближенным. Его следует применять только при достаточно больших значениях $n_{..}$ (например, при $n_{..} > 20$). Ожидаемые значения наблюдений в отдельных клетках не должны быть слишком малыми. Если частотная таблица имеет более чем одну степень свободы, то можно придерживаться правила, сформулированного Кокреном [13], которое состоит в том, что минимальное ожидаемое значение должно быть не меньше единицы. При этом считается выполненным следующее требование: в таблице достаточно мало клеток, ожидаемые значения в которых меньше пяти (под словами «достаточно мало» можно понимать, что таких клеток не более одной, если всего кле-

ток не более пяти, или таких клеток не более двух, если общее число клеток не менее десяти).

2. Для частотных таблиц размера 2X2 существует точный критерий Фишера. Его следует применять вместо критерия (12.9.1) при $n < 20$, а также при $20 \leqslant n < 40$, если наименьшие значения меньше пяти (см. параграф 12.10).

3. При использовании формулы (12.9.1) для таблиц размера 2X2 необходимо применять поправку на непрерывность (Йейтса). Следует обычным образом вычислить ожидаемые числа, но прежде чем подставлять их в (12.9.1), надо вычесть 0,5 из абсолютной величины разности между наблюдаемым числом и ожидаемым.

4. В случаях когда таблица имеет размер 2X2; 2Xs или $r \times 2$, зачастую пользуются специальными вычислительными формулами.

5. Несмотря на то что вычисление значения критериальной статистики совпадает с вычислением в параграфе 12.8, оно основано на другой статистической модели. Как мы отмечали, распределение чисел по клеткам является полиномиальным. Данный критерий проверки зависимости основывается на условном распределении чисел по отдельным строкам или столбцам. Это распределение в случае таблицы 2X2 оказывается гипергеометрическим (см. параграф 10.13).

Пример 12.9.1. В своей статье по генетике [78] Батесон и Паннетт регистрировали цвет и форму 427 частиц пыльцы. Эти данные воспроизведены в табл. 12.9.1. Позволяют ли они предположить, что зависимость между цветом и формой частиц пыльцы существует?

Таблица 12.9.1. **Цвет и форма 427 частиц пыльцы**

Форма	Цвет		Всего
	пурпурные	красные	
	Число частиц		
Вытянутые	296	27	323
Круглые	19	85	104
Всего	315	112	427

Необходимо проверить гипотезы:

H_0: зависимости не существует;

H_1: зависимость существует.

5 %-ная критическая область представляет собой верхнюю 5 %-ную область для распределения χ^2_1, что соответствует значению критериальной статистики, большему, чем 3,84. Если зависимости нет, то ожидаемое число пурпурных вытянутых частиц должно быть равно: 315x323/427 = 238,28, ожидаемое число пурпурных круглых частиц равно: $315 \times 104/427 = 76{,}72$, ожидаемое число красных вытянутых частиц равно: $112 \times 323/427 = 84{,}72$ и ожидаемое число красных круглых частиц равно: $112 \times 104/427 = 27{,}28$. Значение критерия равно:

$$\chi^2 = (296 - 238{,}28 - 0{,}5)^2/238{,}28 + (19 - 76{,}72 + 0{,}5)^2/76{,}72 +$$
$$+ (27 - 84{,}72 + 0{,}5)^2/84{,}72 + (85 - 27{,}28 - 0{,}5)^2/27{,}28 = 215{,}1.$$

Эта величина весьма значима, тем самым мы отклоняем нулевую гипотезу и принимаем альтернативную: есть свидетельство в пользу того, что зависимость между цветом и формой частиц пыльцы существует.

Необходимо сравнить этот пример с примером 12.7.2, в котором учитывалось ожидаемое распределение числа частиц по клеткам.

Литература: [2, с. 52—66; русский перевод с. 77—96], [3, с. 224], [7, с. 211—217; русский перевод с. 198—204], [9, с. 312—315], [45, с. 241-244], [70, с. 311-318], [86, с. 235], [91, с. 217—231], [105, с. 60—69].

12.10. ТОЧНЫЙ КРИТЕРИЙ ФИШЕРА ДЛЯ ТАБЛИЦ СОПРЯЖЕННОСТИ ПРИЗНАКОВ 2X2

Обычная форма записи данных. В параграфе 12.9 рассматривалась таблица сопряженности признаков общего вида размера rXs. Она была описана при помощи двух переменных, которые определяли цвет глаз и волос соответственно. Таблица размера 2x2 имеет следующий вид:

Класс „цвет волос"	Класс „цвет глаз"		Всего
	1	2	
1	ли	"12	"1.
2	"21	"22	"2.
Всего	"Л	$n_{.2}$	$n_{..}$

Статистическая модель. Распределение чисел, стоящих в каждом из четырех клеток, является полиномиальным. Условное распределение значений в клетках строки или столбца гипергеометрическое (см. п. 5 комментариев в параграфе 12.9).

Гипотезы.

H_0: между цветом волос и цветом глаз индивидуума не существует зависимости;

H_1: нулевая гипотеза неверна.

Критическая область. Область, определяемая неравенством $P \leqslant 0{,}05$, где P — значение критерия.

Вычисление значения критериальной статистики. Таблица сопряженности признаков 2X2 с элементами столбца и строки «всего», равными $m.$, $n_{2.}$, $n_{.1}$ и $n_{.2}$, и общим объемом выборки $n..$ полностью определяется заданием числа, стоящего в верхней левой клетке (таблица имеет одну степень свободы). Таким образом, можно найти область допустимых значений для этого числа f_{11}. Максимум и минимум значений f_{11} обозначаются через M и m соответственно. Если выбрано некоторое значение f_{11}, то однозначно определяются f_{12}, f_{21}, f_{22}. Вероятность получения таблицы сопряженности признаков со значением f_{11} в верхнем левом углу при заданных выше элементах столбца и строки «всего» равна:

$$p(f_{11}) = \left(\frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}{n_{..}!}\right) \frac{1}{f_{11}!\, f_{12}!\, f_{21}!\, f_{22}!}. \qquad (12.10.1)$$

Допустим, что $n_{1.} < n_{2.}$ и $n_{.1} < n_{.2}$ (строки и столбцы таблицы всегда могут быть переставлены так, что это требование будет выполнено). Если n_{11} меньше, чем $n_{1.}n_{.1}/n_{..}$, то необходимо действовать по следующей схеме вычислений:

1. Вычислить вероятности событий от n_{11} до m: $p(n_{11})$, $p(n_{11}-1), \ldots, p(m)$.
2. Вычислить вероятности событий от M до M^*: $p(M)$, $p(M-1), \ldots, p(M^*)$. M^* — целое число, для которого $p(M^*) < p(n_{11}) < p(M^*-1)$.
3. Значение критерия равно сумме вычисленных выше значений от $p(n_{11})$ до $p(m)$ и от $p(M)$ до $p(M^*)$.

В случае, когда $n_{11} > n_{1.}n_{.1}/n_{..}$, необходимо действовать так:

1. Вычислить вероятности событий от n_{11} до M: $p(n_{11})$, $p(n_{11}+1), \ldots, p(M)$.
2. Вычислить вероятности событий от m до m^*: $p(m)$, $p(m+1), \ldots, p(m^*)$; m^* — целое число, такое, что $p(m^*) < p(n_{11}) < p(m^*+1)$.
3. Значение критериальной статистики P есть сумма вычисленных выше двух рядов вероятностей.

Комментарии. 1. По возможности следует применять более удобный критерий χ^2 (см. параграф 12.9). В п. 2 комментариев в параграфе 12.9 объясняются условия, при которых критерий χ^2 неприменим и приходится обращаться к точному критерию Фишера.

2. Вместо вычисления отдельных вероятностей можно воспользоваться табличными данными. Таблицы приведены в книге Зара [105, с. 518—542].

Пример 12.10.1. Есть ли основания предполагать, что в следующей таблице сопряженности признаков последние зависимы?

4	2	6
1	8	9
5	10	15

Для проверки гипотезы независимости в данном случае критерий χ^2 не приемлем. Следовательно, необходимо применить точный критерий Фишера. Если элементы строки и столбца «всего» фиксированы и равны значениям в нашей таблице, то наименьшее и наибольшее значения в левом верхнем углу равны $m = 0$ и $M = 5$ соответственно. Значение наблюдения, стоящее в этой клетке, равно 4. Это больше, чем ожидаемое значение, которое равно: $(5\times6)/15 = 2$. Начнем с вычисления

$$\frac{6!\,9!\,5!\,10!}{15!} = 87\,004{,}2.$$

Тогда

$$p(4) = 87\,004{,}2\,\frac{1}{4!\,2!\,1!\,8!} = 0{,}044955,$$

$$p(5) = 87\,004{,}2\,\frac{1}{4!\,1!\,0!\,9!} = 0{,}001\,998,$$

$$p(0) = 87\,004{,}2\,\frac{1}{0!\,6!\,5!\,4!} = 0{,}041\,958,$$

$$p(1) = 87004{,}2\,\frac{1}{1!\,5!\,4!\,5!} = 0{,}251\,748.$$

Значение критериальной статистики равно: $P = p(4) + p(5) + p(0) = 0{,}0889$. Эта величина не является значимой. Поэтому у нас нет свидетельств в пользу гипотезы зависимости признаков.

Литература: [7, с. 163—166; русский перевод с. 151—154], [105, с. 291—295].

12.11. КРИТЕРИЙ χ^2

Предположим, что к некоторым данным было подобрано распределение, имеющее Y параметров. Необходимо проверить пригодность подбора. Изложенный в п. 4 комментариев параграфа 12.7 метод χ^2 (он был продемонстрирован в примере 12.7.3) не подходит для этой цели, так как параметры распределения в данном случае оцениваются на основе данных, а не подбираются заранее. Тем не менее оказывается, что и для этой цели можно применить критерий χ^2. При этом необходимо немного модифицировать его, уменьшив число степеней свободы;

Обычная форма записи данных. Пусть по некоторым данным было построено распределение с γ параметрами. Область возможных значений переменной разделена на k непересекающихся интервалов или клеток. Каждое из n начальных наблюдений должно лежать в одной из этих клеток. Ожидаемое значение в каждой клетке вычисляется на основе подобранного распределения.

	Число наблюдений в клетке				Всего
	1	2	...	k	
Наблюдаемое число	o_1	o_2	...	o_k	n
Ожидаемое число	e_1	e_2	...	e_k	n

Статистическая модель. Наблюдения являются независимыми, вероятность попадания каждого наблюдения в клетку 1 равна p_1, в клетку $2 - p_2$, ..., в клетку $k - p_k$. Тем самым распределение по k клеткам является полиномиальным.

Гипотезы.

$$H_0: p_1 = e_1/n,\ p_2 = e_2/n, \ldots,\ p_k = e_k/n;$$

H_1: нулевая гипотеза неверна.

Критическая область. Область, лежащая выше верхней 5 %-ной точки распределения χ^2 с $k - \gamma - 1$ степенями свободы.

Вычисление значения критериальной статистики.

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \ldots + \frac{(o_k - e_k)^2}{e_k}. \qquad (12.11.1)$$

Комментарии. 1. Ожидаемое число в каждой клетке не должно быть слишком мало. В случае унимодального распределения малые ожидаемые значения будут лишь на одном или на обоих его концах. Кокрен [13] рекомендует группировку клеток, при которой минимальные ожидаемые значения на обоих концах будут не меньше единицы.

2. Если ожидаемое значение в некоторой клетке слишком мало, то необходимо объединить ее с соседней, складывая соответствующие как ожидаемые, так и наблюдаемые значения, а затем уменьшить число степеней свободы па единицу.

3. В случае непрерывного распределения каждая клетка будет соответствовать области возможных значений непрерывной переменной (см. пример 12.7.3).

4. Для проверки качества согласования иногда применяется также метод Колмогорова—Смирнова, описанный в параграфе 12.12. Преимущество этого критерия состоит в том, что принимается во внимание порядок клеток.

5. Заметим, что распределение значений, попадающих в различные клетки, полиномиально, хотя подобранное распределение может быть нормальным, пуассоновским или еще каким-нибудь.

Пример 12.11.1. В примере 10.11.1 к экспериментальным данным, описывающим распределение яиц паразитических нематод у овец, было подобрано отрицательное биномиальное распределение со средним, равным 6,73, и параметром $k = 0{,}6$. Наблюдаемые и ожидаемые значения представлены в табл. 10.11.1. Требуется проверить подбор распределения методом χ^2.

Ожидаемые частоты появления 6, 7, 8 и 9 яиц малы. Поэтому мы объединим клетки 6 и 7, а также клетки 8 и 9. Таким образом, в эксперименте будем рассматривать одиннадцать клеток.

После подгонки отрицательного биномиального распределения к имеющимся данным были вычислены два параметра: k и p. Таким образом, 5 %-ная критическая область для критерия согласия является верхней 5 %-ной областью распределения χ^2_8, что соответствует значению 15,51. Значение критериальной статистики равно:

$$\chi^2 = (20 - 20{,}0)^2/20{,}0 + (12 - 11{,}0)^2/11{,}0 + \ldots$$
$$\ldots + (11 - 7{,}8)^2/7{,}8 = 12{,}1.$$

Эта величина не значима, отсюда мы заключаем, что распределение адекватно данным.

Литература: [2, с. 67—77; русский перевод с. 97—110], [9, с. 390—391], [13], [16, с. 132—134], [70, с. 308—311], [105, с. 41, 45—50].

12.12. КРИТЕРИЙ КОЛМОГОРОВА-СМИРНОВА ДЛЯ ОДНОЙ ВЫБОРКИ

В п. 4 комментариев из параграфа 12.7 был описан метод проверки гипотез с использованием χ^2, он был применен в примере 12.7.3. Критерий Колмогорова—Смирнова является другим непараметрическим критерием, который может быть применен для этой цели; его преимущество перед критерием χ^2 связано с тем, что он принимает во внимание порядок наблюдений. Критерий предназначен для проверки согласия эмпирической и теоретической функций распределений.

Критерий Колмогорова—Смирнова основывается на статистической модели, которая предполагает непрерывность распределения, так что вероятность совпадения выборочных значений равна нулю. Однако на практике критерий часто применяется к сгруппированным данным и данным выборок из дискретных распределений. 15 обоих этих случаях учитывается возможность появления равных значений наблюдений. В данном случае уровень значимости критерия ниже номинального, и вероятность ошибки второго рода возрастает.

Как критерий χ^2 из примера 12.7.3, так и критерий Колмогорова—Смирнова предполагают, что распределение, фигурирующее в нулевой гипотезе, должно быть полностью определено заранее (например, оно может быть нормальным с нулевым средним и единичной дисперсией). При работе с критерием согласия χ^2 необходимо заранее определить тип распределения, а параметры оцениваются по выборочным данным. Мы видели, что критерий χ^2 легко модифицируется при помощи уменьшения числа степеней свободы, но неизвестно, какие изменения должны быть внесены в процедуру применения критерия Колмогорова—Смирнова. Тем не менее иногда он применяется в качестве критерия при проверке гипотез о законе распределения. Следует учитывать, что в этом случае истинный уровень значимости будет несколько ниже номинального и возрастет вероятность ошибок второго рода. По всей видимости, этот эффект не будет велик, если число оцениваемых параметров мало по сравнению с объемом выборки.

Обычная форма записи данных. $x_1, x_2, \ldots, x_n$ — выборка, состоящая из n независимых наблюдений, упорядоченных так, что $x_1 < x_2 < \ldots < x_n$.

Статистическая модель. Наблюдения независимы и берутся из генеральной совокупности, распределение которой предполагается непрерывным.

Гипотезы.

H_0: функция распределения равна $F(x)$;

H_1: нулевая гипотеза неверна.

Критическая область. Область со значениями больше 5 %-ной верхней точки распределения Колмогорова—Смирнова (см. табл. 12.12.1).

Вычисление значения критериальной статистики. 1. Вычислим кумулятивные разности: $D_1 = 1 - nF(x_1)$, $D_2 = 2 - nF(x_2)$ и т. д.

2. Найдем $|D_i|_{\max}$, наибольшее абсолютное значение кумулятивных разностей.

3. Вычислим значение критерия $D = (|D_i|_{\max})/n$.

Комментарий. Следует избегать применения этого критерия в качестве критерия согласия. Вследствие неточностей, возникающих при определении его уровня значимости, следует также избегать его применения при дискретных или непрерывных сгруппированных данных. С другой стороны, можно заметить, что альтернативный критерий χ^2 не всегда отражает накопление большого числа малых отклонений одного знака. Если регистрация такого накопления имеет значение, то допустимо использование критерия Колмогорова—Смирнова. При этом необходимо соблюдать некоторую осторожность. Если значение статистического критерия превосходит номинальное критическое значение, то нулевая гипотеза отклоняется. Если его величина намного меньше номинального критического значения, следует принять нулевую гипотезу. В том случае, когда значение критериальной статистики лишь немного меньше номинального, нужно быть осторожным в принятии определенного решения.

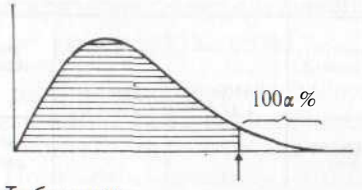
Пример 12.12.1. В табл. 12.7.3 приведена выборка, состоящая из двадцати независимых наблюдений из генеральной совокупности с неизвестным распределением. С помощью метода Колмогорова—Смирнова проверим нулевую гипотезу, предполагающую, что это нормальное распределение с нулевым средним и единичной дисперсией.

Начнем с упорядочивания наблюдений по возрастанию (см. табл. 12.12.2), а затем вычислим наблюдаемые и ожидаемые значения, которые меньше каждого наблюдения или равны ему. Например, зафиксированное число наблюдений, которые не больше $-0{,}97$, равно шести, ожидаемое значение равно: $20\Phi(-0{,}97) = 3{,}32$. Абсолютное значение кумулятивной разности при $x = -0{,}97$ равно, таким образом, 2,68. Аналогично вычисляется абсолютное значение кумулятивной разности в других точках.

Из табл. 12.12.2 видим, что максимальная абсолютная величина разности равна 5,96 и значение критериальной статистики равно:

$$D = 5{,}96/20 = 0{,}298.$$

Таблица 12.12.1. Верхние 100α %-ные точки распределения Колмогорова—Смирнова



n*	a = 0,05	0,01	0,001
5	0,563	0,669	0,781
6	0,519	0,617	0,725
7	0,483	0,576	0,679
8	0,454	0,542	0,641
9	0,430	0,513	0,608
10	0,409	0,489	0,580
11	0,391	0,468	0,556
12	0,375	0,449	0,534
13	0,361	0,432	0,515
14	0,349	0,418	0,498
15	0,338	0,404	0,482
16	0,327	0,392	0,467
17	0,318	0,381	0,454
18	0,309	0,371	0,442
19	0,301	0,361	0,431
20	0,294	0,352	0,421
25	0,264	0,317	0,378
30	0,242	0,290	0,347
35	0,224	0,269	0,322
40	0,210	0,252	0,302
45	0,198	0,238	0,285
50	0,188	0,226	0,271

* При $n>50$ критическое значение D может быть приближенно вычислено как $\left\{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)\right\}^{\frac{1}{2}}$.

Источник. [105]. Воспроизведено с разрешения фирмы Prentice-Hall, Inc., Englewood Cliffs. New Jersey, USA.

5 %-ное критическое значение равно 0,294; поэтому мы отклоняем нулевую гипотезу и заключаем, что неизвестное распределение не является нормальным с нулевым средним и единичной дисперсией. Это заключение противоположно полученному по крите-

Таблица 12.12.2. **Проверка по методу Колмогорова—Смирнова нулевой гипотезы, состоящей в том, что неизвестное распределение нормально с нулевым средним и единичной дисперсией**

Наблюдение X_i	Число наблюдений, меньших или равных x_i		Разность $\lvert D_i \rvert$	Наблюдение X_i	Число наблюдений, меньших или равных x_i		Разность $\lvert D_i \rvert$
	наблюдаемое	ожидаемое			наблюдаемое	ожидаемое	
-2,41	1	0,16	0,84	-0,43	11	6,67	4,33
-1,93	2	0,54	1,46	-0,41	12	6,82	5,18
— 1,82	3	0,69	2,31	-0,38	13	7,04	5,96
-1,23	4	2,19	1,81	-0,21	14	8,34	5,66
-1,22	5	2,22	2,78	0,33	15	12,59	2,41
-0,97	6	3,32	2,68	0,45	16	13,54	2,46
-0,95	7	3,42	3,58	0,48	17	13,69	3,31
-0,83	8	4,07	3,93	1,29	18	18,03	3,03
-0,52	9	6,03	2,97	1,47	19	18,58	0,42
-0,44	10	6,60	3,40	2,66	20	19,92	0,08

рию χ^2, который не учитывает большую кумулятивную разность между наблюдаемой выборкой и выборкой из стандартного нормального распределения.

Следует обратить внимание на предупреждение в параграфе 12.2.

Пример 12.12.2. В примере 10.11.1 по некоторым данным, описывающим распределение яиц паразитических нематод у овец, было подобрано отрицательное биномиальное распределение со средним 6,73 и параметром $k = 0{,}6$. Наблюдаемые и ожидаемые значения представлены в табл. 10.11.1. Проверим согласованность по методу Колмогорова—Смирнова.

Таблица **12.12.3. Кумулятивное распределение яиц паразитических нематод у шотландских овец и кумулятивное ожидаемое распределение, вычисленное на основе отрицательного биномиального распределения (см. пример 10.11.1)**

Число яиц	Кумулятивная частота		Кумулятивная разность	Число яиц	Кумулятивная частота		Кумулятивная разность
	наблюдаемая	ожидаемая			наблюдаемая	ожидаемая	
0	20	20,0	0,0	7	68	62,6	5,4
1	32	31,2	1,0	8	70	65,5	4,5
2	46	39,1	6,9	9	73	68,1	4,9
3	53	45,6	7,4	10—14	77	77,1	—0,1
4	56	50,9	5,1	15-19	79	82,2	—3,2
5	62	55,4	6,6	20+	90	90,0	0,0
6	65	59,3	5,7				

Начнем с построения табл. 12.12.3. Согласно формуле, приведенной в сноске к табл. 12.12.1, 5 %-ное критическое значение равно 0,143. Значение критерия будет следующим:

$$D = 7{,}4/90 = 0{,}082,$$

что намного меньше номинального критического значения. Однако по формуле в сноске к табл. 12.12.1 можно определить, что номинальная вероятность, соответствующая значению $D = 0{,}082$, равна приблизительно 0,6. Поэтому мы считаем, что подобранное распределение адекватно данным.

Литература: [3, с. 225], [9, с. 392-393], [64], [105, с. 54-58].

12.13. КРИТЕРИЙ ПРОВЕРКИ ЗНАЧЕНИЯ СРЕДНЕГО

Обычная форма записи данных. $x_1, x_2, \ldots, x_n$ — выборка объема n.

Статистическая модель. Наблюдения независимы и выбираются из нормальной совокупности со средним μ и дисперсией σ^2.

Гипотезы.

а) *Равенство* $H_0: \mu = \mu_0$; $H_1: \mu \neq \mu_0$.

б) *Неравенство* $H_0: \mu \leqslant \mu_0$; $H_1: \mu > \mu_0$.

в) *Неравенство* $H_0: \mu \geqslant \mu_0$; $H_1: \mu < \mu_0$.

Критическая область.

а) *Равенство.* Значения выше, чем верхняя 2,5 %-ная, или ниже, чем нижняя 2,5 %-ная точка t_{n-1}-распределения.

б) *Неравенство.* Значения выше, чем верхняя 5 %-ная точка t_{n-1}-распределения.

в) *Неравенство.* Значения ниже, чем нижняя 5 %-ная точка t_{n-1}-распределения.

Вычисление значения критериальной статистики. $\bar{x}$ — выборочное среднее (формула (8.4.1)); s^2 — выборочная дисперсия (формула (8.4.3)). Положим $\mu = \mu_0$ в статистике

$$t = n^{\frac{1}{2}} (\bar{x} - \mu)/s. \qquad (12.13.1)$$

Комментарии. 1. μ_0 есть некоторое число, например 103,6.

2. Иногда известна дисперсия генеральной совокупности σ^2. Тогда следует использовать стандартное нормальное распределение вместо t-распределения, при этом нет необходимости вычислять s^2. Стандартное отклонение распределения в данном случае заменяет s в формуле (12.13.1).

3. Критерий устойчив при умеренных отклонениях распределения от нормального. Следует также обратить внимание читателя на непараметрический критерий, описанный в параграфе 12.14.

I. В силу симметрии t-распределения при проверке равенства достаточно сравнивать значение $|t|$ с верхней 2,5 %-ной точкой распределения.

Пример 12.13.1. Статистическому анализу были подвергнуты одиннадцать экземпляров ручной гранаты. Требовалось проверить утверждение их производителя, что среднее время срабатывания взрывателя[2] равно 4,01 с. При проверке получена следующая выборка: 4,21; 4,03; 3,99; 4,05; 3,89; 3,98; 4,01; 3,92; 4,23; 3,85 и 4,20. К какому можно прийти заключению?

Ручная граната будет непригодна для боевого использования, если среднее время срабатывания взрывателя слишком мало (по очевидным причинам) или если оно слишком велико (потому, что тогда противник может схватить гранату и бросить ее обратно). Мы должны проверить гипотезы:

$$H_0: \mu = 4{,}01;$$
$$H_1: \mu \neq 4{,}01.$$

Из таблицы распределения случайной величины t_{10} видно, что критическая область определяется неравенством $|t| > 2{,}228$. Среднее значение x равно 4,033, выборочное стандартное отклонение равно 0,130. Значение критерия есть

$$t = (11)^{1/2}\ (4{,}033 - 4{,}01)/0{,}130 = 0{,}59.$$

Эта величина не является значимой; таким образом, у нас нет оснований сомневаться в утверждении производителя.

Литература: [3, с. 98, 123], [7, с. 105—109, 113—118, 258—260, 295—297; русский перевод с. 94—99, 104—109, 241—244 270—272], [9, с. 323], [39, с. 14—15], [45, с. 170-174], [70, с. 301—302], [86, с. 200—209], [105, с. 86—91].

12.14. НЕПАРАМЕТРИЧЕСКИЙ КРИТЕРИЙ, ОСНОВАННЫЙ НА МЕДИАНЕ

Обычная форма записи данных. $x_1, x_2, x_3, \ldots, x_n$ — выборка объема n.

Статистическая модель. Наблюдения независимы и взяты из одной совокупности.

Гипотезы.

а) *Равенство*	*б) Неравенство*	в) *Неравенство*
H_0: медиана $= m$;	H_0: медиана $\leqslant$ т;	H_0: медиана $\geqslant$ т;
H_1: медиана $\neq m$.	H_1: медиана $> m$.	H_1: медиана $<$ т.

Критическая область.

а) *Равенство.* Выше, чем верхняя 2,5 %-ная, или ниже, чем нижняя 2,5 %-ная, точки стандартного нормального распределения.

б) *Неравенство.* Значения ниже, чем нижняя 5 %-ная точка стандартного нормального распределения.

[2] Колеблемость времени срабатывания, возможно, является еще более важным фактором. Эта проблема обсуждается в примере 12.28.1.

в) *Неравенство.* Значения ниже, чем нижняя 5 %-ная точка стандартного нормального распределения.

Вычисление значения критериальной статистики. Отметим знаком «плюс» выборочное значение, соответствующее данному наблюдению, если оно больше m, и знаком «минус» в противоположном случае. Обозначим через N общее число плюсов. Критериальная статистика равна:

$$T = (2N - n)/\sqrt{n}. \qquad (12.14.1)$$

Комментарии. 1. Медиана распределения определена в параграфе 8.5. В случае симметричного распределения медиана и среднее совпадают и критерий (12.14.1) может применяться для проверки гипотез, относящихся к среднему.

2. Критериальная статистика (12.14.1) может применяться лишь при больших n (скажем, $n > 25$). Для меньших значений n в качестве критериальной статистики надо использовать N, а в качестве критических областей — следующие:

а) *Равенство.* Верхняя 2,5 %-ная или нижняя 2,5 %-ная области биномиального $\left(n, \frac{1}{2}\right)$ распределения.

б) *Неравенство.* Верхняя 5 %-ная область биномиального $\left(n, \frac{1}{2}\right)$ распределения.

в) *Неравенство.* Нижняя 5 %-ная область биномиального $\left(n, \frac{1}{2}\right)$ распределения.

Верхние и нижние 100α %-ные точки биномиального $\left(n, \frac{1}{2}\right)$ распределения приведены в табл. 12.23.1.

3. Если n мало, мощность критерия не очень велика.

Пример 12.14.1. Для определения неизвестного среднего значения совокупности, распределение которой нормально и имеет г...ную дисперсию, была произведена выборка. Объем выборки равен 35, при этом значения двадцати пяти наблюдений оказались больше 10, а значения десяти наблюдений меньше 10. Имеет ли основания предположение о том, что среднее равно 10?

Нормальное распределение симметрично; следовательно, для проверки нулевой гипотезы, состоящей в том, что величина среднего равна 10, можно применить критерий медианы. Для получения 5 %-ного уровня значимости в качестве критической области необходимо использовать объединение верхней и нижней 2,5 %-ных об...П стандартного нормального распределения. Эта область определяется неравенством $|T| > 1{,}96$. Значение критериальной статистики равно:

$$T = (2 \times 25 - 35)/\sqrt{35} = 2{,}54.$$

Это значение является значимым. Поэтому следует отклонить нулевую гипотезу и прийти к выводу, что среднее значение не равно 10.

Литература: [3, с. 222—223].

12.15. РАВЕНСТВО (НЕРАВЕНСТВО) ДВУХ СРЕДНИХ — СЛУЧАЙ РАВНЫХ ДИСПЕРСИЙ

Обычная форма записи данных.

Выборка 1 (объема n_1)	Выборка 2 (объема n_2)
x_{11}	x_{12}
x_{21}	x_{22}
.	.
.	.
.	.
$x_{n_1 1}$	$x_{n_2 2}$
Всего $T_{\cdot 1}$	$T_{\cdot 2}$

Статистическая модель. Обе выборки извлечены из совокупностей, имеющих нормальные распределения с дисперсиями, равными σ^2. Среднее первой совокупности равно μ_1, второй — μ_2. Все наблюдения независимы.

Гипотезы.

а) *Равенство*

$H_0: \mu_1 = \mu_2;$

$H_1: \mu_1 \neq \mu_2.$

б) *Неравенство*

$H_0: \mu_1 \leqslant \mu_2;$

$H_1: \mu_1 > \mu_2.$

Критическая область.

а) *Равенство.* Значения выше, чем верхняя 2,5 %-ная, или ниже, чем нижняя 2,5 %-ная, точки t-распределения с $n_1 + n_2 - 2$ степенями свободы.

б) *Неравенство.* Значения выше, чем верхняя 5 %-ная точка t-распределения с $n_1 + n_2 - 2$ степенями свободы.

Вычисление значения критериальной статистики.

$$x_{\cdot 1} = T_{\cdot 1}/n_1;$$

$$x_{\cdot 2} = T_{\cdot 2}/n_2;$$

s_1^2 — выборочная дисперсия для первой выборки (формула (8.4.3));

s_2^2 — выборочная дисперсия для второй выборки.

Положим $\mu_1 = \mu_2$ в формуле

$$t = \frac{\{(\bar{x}_{\cdot 1} - \bar{x}_{\cdot 2}) - (\mu_1 - \mu_2)\}\,(n_1 + n_2 - 2)^{\frac{1}{2}}}{\{(1/n_1) + (1/n_2)\}^{\frac{1}{2}}\,\{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2\}^{\frac{1}{2}}}. \qquad (12.15.1)$$

Комментарии. 1. Проверка равенства а) эквивалентна дисперсионному анализу по одному признаку для двух выборок (см. параграф 12.18).

2. Критерий устойчив, если распределения совокупностей умеренно отклоняются от нормального. Следует обратить внимание

также на непараметрический критерий, рассмотренный в параграфе 12.17.

3. Если имеется умеренное отклонение от выполнения требования о равенстве дисперсий, то критерий также является устойчивым, когда n_1 и n_2 приблизительно равны. В параграфе 12.16 описан приближенный критерий, не предполагающий равенства дисперсий. В параграфе 12.29 рассматривается критерий проверки равенства дисперсий.

4. Можно проверить и другие гипотезы, например:

в) H_0: $\mu_1 - \mu_2 = \delta$ (здесь δ — некоторая постоянная);

H_1: $\mu_1 - \mu_2 \neq \delta$.

Применим метод а) и положим $(\mu_1 - \mu_2)$ равным δ в формуле **(12.15.1).**

г) $H_0: \mu_1 - \mu_2 \leqslant \delta;$

$H_1: \mu_1 - \mu_2 > \delta.$

Применим метод б) и положим $(\mu_1 - \mu_2)$ равным δ в формуле (12.15.1).

5. Для значений n_1, n_2, таких, что $(n_1 + n_2 - 2)$ больше 30, вместо t-распределения можно воспользоваться нормальным распределением (см. параграф 9.3).

6. В случае когда наблюдения объединены в пары, следует применять метод, рассмотренный в параграфе 12.22.

7. В тех случаях, когда известна общая дисперсия генеральных совокупностей, необходимо использовать σ^2 вместо s_1^2 и s_2^2 и таблицу нормального распределения вместо таблицы t-распределения.

8. Для проверки гипотезы о равенстве средних в силу симметрии t-распределения достаточно сравнивать $|t|$ с верхней 2,5 %-ной точкой.

Пример 12.15.1. Сравниваются по урожайности два сорта пшеницы. Сорт A — обычная разновидность, а сорт B — новый гибрид. Было засеяно двадцать пять акров пшеницей каждого сорта, причем условия созревания на обоих участках были одинаковы. Средний урожай сорта A — 32,0 бушеля на акр с дисперсией, равной 5,9. Средний урожай сорта B — 36,2 бушеля с дисперсией, равной 11,2. Является ли урожайность сорта B значительно более высокой, чем урожайность сорта Л?

Необходимо проверить:

$$H_0: \mu_A \geqslant \mu_B;$$

$$H_1: \mu_A < \mu_B.$$

Примем 5 %-ный уровень значимости. Критическая область есть нижняя 5 %-ная область t_{48}-распределения, что соответствует значению $c = -1{,}645$. Значение критериальной статистики равно:

$$t = \frac{(32{,}0 - 36{,}2)\,48^{\frac{1}{2}}}{\left(\frac{1}{25} + \frac{1}{25}\right)^{\frac{1}{2}}\,(24 \times 5{,}9 + 24 \times 11{,}2)^{\frac{1}{2}}} = -5{,}08.$$

Эта величина (в высокой степени) значима, и мы приходим к выводу, что сорт *B* дает больший урожай.

Пример 12.15.2. Данные в табл. 12.15.1 показывают уровень сывороточного холестерола у семи самцов и шести самок черепахи. Верно ли, что самцы и самки имеют одинаковый уровень концентрации холестерола?

Таблица 12.15.1. **Уровень сывороточного холестерола у черепах**

Самцы	Самки
226,5	221,5
224,1	230,2
218,6	223,4
220,1	224,3
228,8	230,8
229,6	223,8
222,5	

Мы должны проверить:

$$H_0: \mu_M = \mu_F;$$
$$H_1: \mu_M \neq \mu_F.$$

Примем 5 %-ный уровень значимости. Критическая область образуется объединением верхней и нижней 2,5 %-ных областей t_{11}-распределения (что соответствует $|t|>2,201$). На основании данных вычисляем $\bar{x}_M = 224,314$; $x_F = 225,667$; $s^2_{M} = 17,765$ и $A = 14,951$. Значение критериальной статистики равно:

$$|t| = \left| \frac{(224,314 - 225,667)\ 11^{\frac{1}{2}}}{\left(\frac{1}{7} + \frac{1}{6}\right)^{\frac{1}{2}} (6 \times 17,765 + 5 \times 14,951)^{\frac{1}{2}}} \right| = 0,599.$$

Эта величина не является значимой. Поэтому мы принимаем нулевую гипотезу, предполагающую, что самцы и самки черепахи имеют одинаковый уровень концентрации холестерола.

Литература: [3, с. 126—129], [7, с. 119, 297—299; русский перевод с. 111, 272—275], [9, с. 324], [16, с. 148—151], [39, с. 22—23], [70, с. 303—306], [86, с. 210—215], [91, с. 85—95; русский перевод с. 96—106], [105, с. 105—106].

12.16. РАВЕНСТВО (НЕРАВЕНСТВО) ДВУХ СРЕДНИХ — СЛУЧАЙ НЕРАВНЫХ ДИСПЕРСИЙ

Обычная форма записи данных. См. параграф 12.15.

Статистическая модель. Первая выборка производится из нормальной совокупности со средним μ_1 и дисперсией σ_1^2. Вторая выборка — из нормальной совокупности со средним μ_2 и дисперсией σ_2^2. Все наблюдения независимы.

Гипотезы.

а) *Равенство*	б) *Неравенство*
$H_0: \mu_1 = \mu_2;$	$H_0: \mu_1 < \mu_2;$
$H_1: \mu_1 \neq \mu_2.$	$H_1: \mu_1 > \mu_2.$

Критическая область.

а) *Равенство.* Выше верхней 2,5 %-ной и ниже нижней 2,5 %-ной областей t_ν-распределения, где

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)} - 2. \qquad (12.16.1)$$

б) *Неравенство.* Верхняя 5 %-ная область t_ν-распределения.

Вычисление значения критериальной статистики.

$$\bar{x}_{.1} = T_{.1}/n_1,$$
$$\bar{x}_{.2} = T_{.2}/n_2;$$

s_1^2 — выборочная дисперсия первой выборки (формула (8.4.3));

s_2^2 — выборочная дисперсия второй выборки.

Положим $\mu_1 = \mu_2$ в формуле

$$t = \frac{(\bar{x}_{.1} - \bar{x}_{.2}) - (\mu_1 - \mu_2)}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{\frac{1}{2}}}. \qquad (12.16.2)$$

Комментарии. 1. Этот критерий является приближенным. Если нет оснований предполагать, что дисперсии не равны, следует применить точный критерий, описанный в параграфе 12.15. Метод проверки равенства дисперсий рассматривается в параграфе 12.29.

2. Критерий устойчив при умеренных отклонениях распределений совокупностей от нормальности. Следует обратить внимание на непараметрический критерий, описанный в параграфе 12.17.

3. Желательно избегать вычисления величины ν. Всегда целесообразно делать оценку значения критериальной статистики перед вычислением ν. Во многих случаях и без этого вычисления ясно, является ли величина t_ν значимой. Если сумма (n_1+n_2) больше 30, то можно использовать нормальное распределение вместо t_ν-распределения.

4. Данный метод непригоден в случае парных наблюдений (см. параграфы 12.22—12.24).

5. Могут проверяться другие гипотезы, например:

в) $H_0: \mu_1 - \mu_2 = 6;$

$H_1: \mu_1 - \mu_2 \neq \delta.$

Надо применить метод а) и положить разность $(\mu_1 - \mu_2)$ равной б в формуле (12.16.2).

г) $H_0: \mu_1 - \mu_2 \leqslant$ б;

$H_1: \mu_1 - \mu_2 > \delta.$

Надо применить метод *б)* и положить разность $(\mu_1 - \mu_2)$ равной б в формуле (12.16.2).

6. Если известны генеральные дисперсии σ_1^2 и σ_2^2, то вместо выборочных дисперсий следует использовать их. Вместо таблицы t-распределения в данном случае пригодна таблица нормального распределения, при этом критерий будет точным, а не приближенным.

7. При проверке гипотезы равенства средних в силу симметрии распределения достаточно сравнивать значение $|t|$ с верхней 2,5 %-ной точкой t-распределения.

Пример 12.16.1. В примере 12.15.1 описаны экспериментальные данные, полученные при сравнении урожайности двух сортов пшеницы. Является ли урожайность сорта *B* значительно более высокой, чем урожайность сорта Л?

Есть основания предполагать, что урожайность сорта *B* менее постоянна, чем урожайность сорта *A*. Поэтому может возникнуть сомнение относительно возможности применения критерия, описанного в параграфе 12.15. Можно воспользоваться приближенным методом, изложенным в данном параграфе.

Необходимо проверить:

$H_0: \mu_A \geqslant \mu_B;$

$H_1: \mu_A < \mu_B.$

Снова выбираем 5 %-ный уровень значимости. Поскольку значение суммы (n_1+n_2) велико, критической можно считать 5 %-ную область нормального распределения. Она определяется неравенством $t < -1{,}645$. Значение критериальной статистики есть

$$t = (32{,}0 - \mathbf{36{,}2})/(\mathbf{5{,}9/25} + 11{,}2/25)^{1/2} = -5{,}08.$$

Эта величина является (весьма) значимой. Мы отклоняем нулевую гипотезу и приходим к выводу, что сорт *B* должен давать более высокий урожай.

В иллюстративных целях к этим данным был применен второй критерий. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [7, с. 299—304; русский перевод с. 275—280], [39, с. 22—23], [45, с. 171—175], [91, с. 97—100; русский перевод с. 106—109].

12.17. КРИТЕРИЙ МАННА—УИТНИ ДЛЯ ДВУХ НЕЗАВИСИМЫХ ВЫБОРОК

Когда не выполнены условия, при которых возможно применение обычного t-критерия (например, совокупности, из которых производятся выборки, не нормальны), часто обращаются к непараметрическому критерию, описанному в данном параграфе. Он предназначен для проверки гипотезы о неравенстве двух средних. Этот критерий следует применять также в случае ранжированных данных (т. е. если данные заданы как ранги, например ранжирование по росту в группе из $n = n\backslash + n_2$ студентов, $n\backslash$ из которых юноши, а n_2 — девушки).

Заметим, что рассматриваемая нулевая гипотеза, проверяемая при помощи критерия, состоит в том, что совокупности, из которых производятся выборки, одинаково распределены. Если величина критериальной статистики значима, то мы приходим к выводу, что распределения различны, но это еще не дает возможности заключить, что их средние или медианы не равны. Для такого вывода необходимо предположить, что рассматриваемые распределения идентичны во всех остальных аспектах, например что их дисперсии равны. На практике допустимы умеренные различия в значениях дисперсий, так как критерий малочувствителен к ним.

Обычная форма записи данных.

Выборка 1 (объема n_1)		Выборка 2 (объема n_2)	
наблюдение	ранг	наблюдение	ранг
$x\backslash\backslash$	r_{11}	x_{12}	r_{12}
x_{21}	r_{21}	x_{22}	r_{22}
.	.	.	.
.	.	.	.
.	.	.	.
$x_{n_1 1}$	$r_{n_1 1}$	$x_{n_2 2}$	$r_{n_2 2}$
Всего—	R_1	—	R_2

Величины $\{x_{ij}\}$ являются реальными наблюдениями. Определим $n = n_1+n_2$. Числа $\{r_{ij}\}$ — ранги этих наблюдений, они меняются от 1 до n.

Статистическая модель. Все наблюдения независимы; наблюдения, входящие в одну выборку, относятся к одной совокупности.

Гипотезы.

H_0: совокупности одинаково распределены;

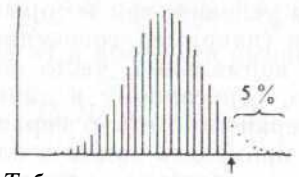
/Л: нулевая гипотеза неверна.

Критическая область.

1) *Малые выборки.* Верхняя 2,5 %-ная область распределения Манна—Уитни с параметрами $n\backslash$ и n_2 (см. табл. 12.17.1).

2) *Большие выборки.* Верхняя 2,5 %-ная область стандартного **Нор**мального распределения.

Таблица 12.17.1. **Верхние 5 %-ные (светлый шрифт) и 2,5 %-ные (жирный шрифт) точки распределения Манна—Уитни* (критическая область включает табличное значение критерия)**



Табличное значение критерия

n_2 \ n_1	2	3	4	5	7	10	12	15	17	20
4	—	12	15							
	—	—	**16**							
5	10	14	18	21						
	—	15	19	23						
6	12	**16**	21	25						
	—	17	22	27						
7	14	19	24	29	38					
	—	20	25	30	41					
8	15	21	27	32	43					
	16	22	28	34	46					
9	**17**	23	30	36	48					
	18	25	32	38	51					
10	19	26	33	39	53	73				
	20	27	35	42	56	77				
12	22	31	39	47	63	86	102			
	23	32	41	49	66	91	107			
14	25	35	45	54	72	99	117	Л 44		
	27	37	47	57	76	104	123	151		
16	29	40	50	61	82	112	132	163	183	
	31	42	53	65	86	118	139	170	191	
18	32	45	56	68	91	125	148	**182**	204	
	34	•47	**60**	72	96	132	155	190	213	
20	36	49	62	75	101	138	163	200	225	262
	38	52	66	80	106	145	171	210	235	273
25	44	**61**	77	93	125	171	202	247	278	323
	47	65	82	98	131	179	**211**	258	290	337
30	53	73	92	111	149	204	240	294	330	384
	55	77	97	117	156	213	251	307	344	400
35	61	84	107	129	172	236	279	341	383	445
	64	89	113	136	181	247	291	356	399	463
40	69	96	121	147	196	269	317	388	435	506
	73	102	129	155	206	281	331	404	453	526

* **Если необходимо, то можно вычислить нижние 100α %-ные точки распределения Манна—Уитни, вычитая верхние ЮОа %-ные точки из n_1n_2.**

Источник. [68]. Воспроизведено с разрешения издателя.

Вычисление значения критериальной статистики.

1) *Малые выборки:*

$$U_1 = n_1n_2 + \frac{1}{2}n_1(n_1+1) - R_1,$$

$$U_2 = n_1n_2 + \frac{1}{2}n_2(n_2+1) - R_2,$$

$$U = \max(U_1,\ U_2).$$

Нужно сравнить U с критическим значением для малых выборок.

2) *Большие выборки.* Следует сравнить с критическим значением для больших выборок статистику

$$T = \left(U - \frac{1}{2}n_1n_2\right)/\{n_1n_2(n+1)/12\}^{\frac{1}{2}}. \qquad (12.17.1)$$

Комментарии. 1. Не имеет значения, как ранжируются наблюдения: от наименьшего до наибольшего или наоборот.

2. Если два или более наблюдения имеют в точности одинаковые значения, они называются *совпадающими. В* этом случае каждому из них следует приписать значение ранга, равное среднему из рангов, которые были бы им приписаны при отсутствии совпадения. Для учета совпадающих наблюдений иногда производится небольшое изменение процедуры вычисления критериальной статистики [105, с. 112—113].

3. Заметим, что $U_1 + U_2 = n_1n_2$. Поэтому если U_1 принадлежит нижней 2,5 %-ной области, то U_2 принадлежит верхней 2,5 %-ной области, следовательно, $U = \max(U_1,\ Uz)$ также принадлежит верхней 2,5 %-ной области. Аналогичное явление имеет место при замене U_1 на Uz и U_2 на U_1. Следовательно, данный критерий двусторонний с 5 %-ным уровнем значимости, что на первый взгляд не очевидно.

4. Могут проверяться другие гипотезы. Пусть выборки производятся из совокупностей, распределение которых совпадает по всем параметрам, за возможным исключением средних значений. Можно проверить следующие гипотезы:

а) $H_0: \mu_1 - \mu_2 = \delta$ (является обусловленной константой);
$H_1: \mu_1 - \mu_2 \neq \delta.$

Вычтем 6 из каждого наблюдения в первой выборке, ранжируем остатки и применим изложенный выше метод.

б) $H_0: \mu_1 - \mu_2 \leqslant \delta;$
$H_1: \mu_1 - \mu_2 > \delta.$

Вычтем δ из каждого наблюдения в первой выборке и затем ранжируем остатки обычным путем. Если значения ранжированы так, что большим значениям приписаны большие (меньшие) ранги и R_1 получается меньше (больше) R_2, то мы автоматически принимаем нулевую гипотезу. В противном случае следует проверить принадлежность U (или T) верхней 5 %-ной критической области.

5. Данные в табл. 12.17.1 могут быть интерполированы. Более обширные таблицы приведены в книге Зара [105, с. 475—487].

6. Табл. 12.17.1 предполагает, что $n_1 \leqslant n_2$. Это не должно вызывать трудностей, так как критические точки, соответствующие параметрам n_1 и n_2, те же, что для n_2 и n_1.

7. Не следует применять критерий к парным наблюдениям (см. параграф 12.24).

8. Если к данным применима статистическая модель из параграфа 12.15, то следует использовать более мощный t-критерий.

Пример 12.17.1. Данные в табл. 12.15.1 иллюстрируют содержание сывороточного холестерола у семи самцов и шести самок черепахи. Верно ли, что самцы и самки черепах имеют одинаковые средние концентрации холестерола?

Допустим, что при возможном различии в средних значениях рассматриваемые распределения в остальном одинаковы. Можно применить метод Манна—Уитни для проверки *.

$$H_0: \mu_M = \mu_F;$$
$$H_1: \mu_M \neq \mu_F.$$

Для уровня значимости в 5 % критическая область определяется неравенством $U \geqslant 36$ (интерполяция показателей табл. 12.17.1). Ранжируя от большего к меньшему, получаем, что сумма рангов для самцов равна 54, а для самок — 37. Отсюда $U_1 = 16$, $U_2 = 26$ и $U = 26$. Эта величина не является значимой. Мы принимаем нулевую гипотезу о том, что средние равны.

В иллюстративных целях к этим данным был применен второй по счету критерий **. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [3, с. 227—228], [16, с. 143—147], [105, с. 109—114].

12.18. ДИСПЕРСИОННЫЙ АНАЛИЗ ПО ОДНОМУ ПРИЗНАКУ ДЛЯ ПРОВЕРКИ РАВЕНСТВА НЕСКОЛЬКИХ СРЕДНИХ

В этом параграфе (а также в параграфе 12.25) мы рассмотрим пример применения весьма общего метода, известного как дисперсионный анализ, который (вопреки своему наименованию) применяется для получения критериев при сравнении средних значений. Общая процедура состоит в том, чтобы определить, какая часть вариаций в экспериментальных результатах вызывается различием в совокупностях, а что может быть отнесено на счет случайных отклонений. Мы имеем возможность оценить важность различий между совокупностями путем сравнения вкладов в дисперсию, которые дают оба эти источника вариации.

* μ_F — среднее значение для самок, μ_M — среднее значение для самцов.— *Примеч. пер.*

** См. пример 12.15.2.— *Примеч. пер.*

Обычная форма записи данных.

Выборка 1 (объема n_1)	Выборка 2 (объема n_2)	...	Выборка k (объема n_k)
x_{11}	x_{12}		x_{1k}
x_{21}	x_{22}		x_{2k}
.	.		.
.	.		.
.	.		.
$x_{n_1 1}$			$x_{n_k k}$
	$x_{n_2 2}$		
Всего $T_{\cdot 1}$	$T_{\cdot 2}$	...	$T_{\cdot k}$

$m.. = T_{\cdot 1} + T_{\cdot 2} + \ldots + T_{\cdot k}; n = n_1 + n_2 + \cdot\ldots + n_k.$

Статистическая модель. Выборки производятся из нормальных совокупностей с дисперсиями, равными σ^2. Первая выборка производится из совокупности со средним μ_1, вторая — со средним μ_2, ..., k-я — из совокупности со средним μ_k. Все наблюдения независимы.

Гипотезы.

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k;$$

H_1: не все средние равны.

Критическая область. Верхняя 5 %-ная область $F_{k-1,\, n-k}$-распределения.

Вычисление значения критериальной статистики. Возведем в квадрат значения всех наблюдений и просуммируем их. Получим $\sum\sum x_{ij}$. Вычислим общую сумму квадратов $\sum\sum x_{ij}^2 - T_{..}^2/n$. Найдем сумму квадратов «между выборками»:

$$(T_{\cdot 1}^2/n_1 + \ldots + T_{\cdot k}^2/n_k) - T_{..}^2/n.$$

Заполним таблицу дисперсионного анализа 12.18.1[3]. Критериальная статистика равна:

$$F = \frac{\text{средний квадрат между выборками}}{\text{остаточный средний квадрат}}. \qquad (12.18.1)$$

Комментарии. 1. В случае двух выборок критерий эквивалентен критерию а) из параграфа 12.15.

2. Критерий устойчив, если распределения рассматриваемых совокупностей умеренно отклоняются от нормальных, при условии, что выборки имеют достаточно большие объемы.

[3] В этой таблице (как и в других таблицах дисперсионного анализа) с. к. означает сумму квадратов, с. с.— степени свободы, ср. к.— средний квадрат.

Таблица 12.18.1. Дисперсионный анализ по одному признаку

Компонента дисперсии* (1)	с. к. (2)	с. с. (3)	ср. к. (4) = (2)/(3)
Между выборками	$(\sum T_{.j}^2/n_j) - T_{..}^2/n$	$k-1$	(определяется делением)
Остаточная	(определяется вычитанием)	$n-k$	
Полная	$\sum\sum x_{ij}^2 - T_{..}^2/n$	$n-1$	—

* Иногда в подобных таблицах эту графу называют «источник вариации».— *Примеч. пер.*

3. Если имеется умеренное отклонение от выполнения требования равенства дисперсий, критерий также устойчив, но только при условии приблизительного равенства объемов выборок. Критерии проверки равенства дисперсий рассмотрены в параграфах 12.29—12.30.

4. В том случае, когда возможность применения критерия к некоторым данным может вызвать сомнение, следует воспользоваться непараметрическим критерием из параграфа 12.19.

5. Можно проверять другие гипотезы, например:

$$H_0: \mu_1 - \mu_2 = 6,\ \mu_2 = \mu_3 = \ldots = \mu_k;$$

H_1: нулевая гипотеза неверна.

В этих случаях следует вычесть δ из значений наблюдений в выборке 1, а затем применить описанную выше процедуру.

6. Если проверка дает значимый результат, то мы приходим к выводу, что не все средние равны. Для определения того, какие средние различаются, следует применить метод множественного сравнения (см. параграф 12.21). Недопустимо проводить попарное сравнение столбцов с помощью t-критерия или изложенной схемы дисперсионного анализа.

7. Указание для вычислительной процедуры. Можно все элементы таблицы умножить (или разделить) на единую константу. Можно также вычесть (или прибавить) единую константу. Конечное значение F от этого не изменится.

8. При вычислении суммы квадратов между выборками соответствующие суммы возводятся в квадрат. Полученная сумма квадратов всегда делится на число элементов этой суммы.

Пример 12.18.1. Биолог производит классификацию червей по трем группам с помощью структурных характеристик. Из каждой группы производится случайная выборка (см. табл. 12.18.2). Верно ли, что средняя длина червя одинакова для каждой группы?

Примем 5 %-ный уровень значимости. Критическая область представляет собой верхнюю 5 %-ную область $F_{2,15}$-распределения. Эта область определяется неравенством $F > 3{,}68$.

Таблица 12.18.2. Длина червей в трех различных группах, см

Группа 1	Группа 2	Группа 3
10,2	12,2	9,2
8,2	10,6	10,5
8,9	9,9	9,2
8,0	13,0	8,7
8,3	8,1	9,0
8,0	10,8	
	11,5	
Всего 51,6	76,1	46,6

Из таблицы 12.18.2 находим:

$$T_1 = 51{,}6 \qquad n_1 = 6$$
$$T_{.2} = 76{,}1 \qquad n_2 = 7$$
$$T_{.3} = 46{,}6 \qquad n_3 = 5$$
$$T_{..} = 174{,}3 \qquad n = 18$$

Вычисляем:

$$\sum\sum x_{ij}^2 = (10{,}2)^2 + (8{,}2)^2 + \ldots + (8{,}7)^2 + (9{,}0)^2 = 1726{,}31.$$

Общая сумма квадратов будет следующей:

$$1726{,}31 - (174{,}3)^2/18 = 38{,}505.$$

Находим сумму квадратов между выборками:

$$(51{,}6)^2/6 + (76{,}1)^2/7 + (46{,}6)^2/5 - (174{,}3)^2/18 = 17{,}583.$$

Теперь можно заполнить таблицу (см. табл. 12.18.3) дисперсионного анализа. Значение критериальной статистики равно:

$$F = 8{,}792/1{,}395 = 6{,}3.$$

Величина F является значимой. Нулевую гипотезу, состоящую в том, что длина червей совпадает во всех трех группах, следует отклонить. Нам неизвестно, в каких группах средние длины червей действительно различаются, хотя мы и пришли к заключению, что не все они одинаковы. Для того чтобы это выяснить, необходимо прибегнуть к методу множественного сравнения (см. параграф 12.21).

Таблица 12.18.3. Дисперсионный анализ по одному признаку

Компонента дисперсии	с. к.	с. с.	ср. к.
Между выборками	17,583	2	8,792
Остаточная	20,922	15	1,395
Полная	38,505	17	—

Литература: [3, с. 161 — 165], [7, с. 309—333; русский перевод с. 283—307], [9, с. 342], [16, с. 171 — 190], [39, с. 26—64], [45, с. 250—256], [70, с. 372—373], [86, с. 282—287], [88, с. 331—368; русский перевод с. 256—286], [91, с. 237—253; русский перевод с. 227—245], [105, с. 133—139].

12.19. НЕПАРАМЕТРИЧЕСКИЙ ДИСПЕРСИОННЫЙ АНАЛИЗ ПО ОДНОМУ ПРИЗНАКУ С ПРИМЕНЕНИЕМ КРИТЕРИЯ КРАСКАЛА—УОЛЛИСА ДЛЯ НЕСКОЛЬКИХ НЕЗАВИСИМЫХ ВЫБОРОК

Если не выполняются предположения, на которых основывается обычный дисперсионный анализ по одному признаку (см. параграф 12.18), то для проверки совпадения нескольких средних часто применяется непараметрический критерий, описанный в данном параграфе. Его можно использовать, когда рассматриваемые совокупности не являются нормально распределенными, а также когда данные представлены в виде рангов (например, самый высокий ребенок, второй по росту ребенок и т. д.).

Заметим, что нулевая гипотеза, проверяемая с помощью критерия, состоит в том, что совокупности, из которых производятся выборки, являются одинаково распределенными. Строго говоря, если критерий дает значимый результат, то можно только утверждать, что распределения совокупностей различны. Это, однако, не означает, что их средние не равны между собой. Для вывода, что выборки производились из совокупности с различными средними (или медианами), необходимо предположить, что совокупности одинаковы по всем другим параметрам, в частности их дисперсии должны быть равны. На практике допустимы, однако, умеренные отклонения от этого правила, поскольку критерий в достаточной мере нечувствителен к ним.

Обычная форма записи данных. См. параграф 12.18 (имеется k выборок, общее число наблюдений равно n).

Статистическая модель. Имеется k совокупностей: каждая выборка извлекается из своей совокупности. Все наблюдения независимы.

Гипотезы.

H_0: все k совокупностей одинаково распределены;

H_1: нулевая гипотеза неверна.

Критическая область.

а) *Малые выборки.* Верхняя 5 %-ная область распределения Краскала—Уоллиса (см. табл. 12.19.1).

б) *Большие выборки.* Верхняя 5 %-ная область распределения χ^2_{k-1}.

Вычисление значения критериальной статистики. Все n наблюдений упорядочены по возрастанию от 1 до n. Находим сумму рангов $R_1, R_2, \ldots, R_k$ для k групп. Вычисляем критерий:

$$H = \frac{12}{n(n+1)}\left(R_1^2/n_1 + \ldots + R_k^2/n_k\right) - 3(n+1). \qquad (12.19.1)$$

Комментарии. 1. Статистический критерий применяется как для больших, так и для малых выборок.

2. Когда значения двух или более наблюдений в точности равны, они называются *совпадающими.* Приписываемый им ранг следует вычислять как среднее из рангов, которые они получили бы при отсутствии совпадений. Для учета совпадающих наблюдений иногда производится небольшое уточнение процедуры вычисления Я (см. [105, с. 142]).

3. Если H_0 отклоняется, то на основе изложенной процедуры нельзя решить, какие совокупности имеют различные распределения. Для сравнения всевозможных пар недопустимо применять критерий Манна—Уитни. Для решения вопроса о том, какие совокупности различны, должен применяться метод множественного сравнения (см. [105, с. 156— 157]).

4. Могут проверяться другие гипотезы. Например:

$$H_0: \mu_1 - 6 = \mu_2 = \ldots = \mu_k;$$

H_1: нулевая гипотеза неверна.

(Здесь δ — обусловленная константа.) Следует вычесть δ из значений всех наблюдений первой выборки, ранжировать их и действовать, как ранее.

5. Недопустимо применять критерий для зависимых наблюдений (см. параграф 12.26).

6. Если к рассматриваемым данным применима статистическая модель из параграфа 12.18, то следует воспользоваться более мощным F-критерием.

Пример 12.19.1. Биолог с помощью структурных характеристик классифицирует червей по трем группам. Из каждой группы производится случайная выборка (см. табл. 12.18.2). Следует ли прийти к заключению, что средняя длина червей одинакова для всех групп?

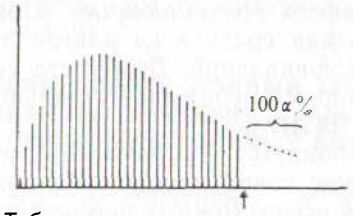
Эта нулевая гипотеза проверялась в параграфе 12.18 посредством дисперсионного анализа по одному признаку. Продемонстрируем применение критерия Краскала—Уоллиса на материале этих данных. Значения n_1, n_2 и n_3 превышают указанные в табл. 12.19.1, поэтому в качестве нашей 5 %-ной критической области мы возьмем верхнюю 5 %-ную область распределения χ^2_2, что соответствует значению критерия, превышающему 5,99.

Когда 18 наблюдений упорядочены по возрастанию от 1 до 18, суммы рангов для трех групп равны: $R_1 = 31$, $R_2 = 94$ и $R_3 = 46$. Значение критериальной статистики следующее:

$$H = \frac{12}{18 \times 19}\left(\frac{31^2}{6} + \frac{942}{7} + \frac{46^2}{5}\right) - 3 \times 19 = 7{,}76.$$

Полученный результат значим, поэтому нельзя предполагать, что три выборки извлечены из одинаково распределенных совокупностей. Если допустить, что совокупности одинаковы по всем параметрам, за возможным исключением средних, то можно заключить, что средние значения совокупностей не совпадают.

Таблица 12.19.1. **Верхние 100α %-ные точки распределения Краскала—Уоллиса (критическая область включает табличное значение критерия)**



n_1	n_2	n_3	a = 0,10	0,05	0,01	n_1	n_2	n_3	a = 0,10	0,05	0,01
2	2	2	4,57	—	—	5	3	1	4,01	4,96	
3	2	2	4,50	4,71	—	5	3	2	4,65	5,25	6,82
3	3	2	4,55	5,36	—	5	3	3	4,53	5,34	6,98
3	3	3	4,62	5,60	7,20	5	4	1	3,98	4,98	6,95
4	2	2	4,37	5,33	—	5	4	2	4,54	5,27	7,11
4	3	2	4,51	5,44	6,44	5	4	3	4,54	5,63	7,44
4	3	3	4,70	5,72	6,74	5	4	4	4,61	5,61	7,76
4	4	1	4,16	4,96	6,66	5	5	1	4,10	5,12	7,30
4	4	2	4,55	5,45	7,03	5	5	2	4,50	5,33	7,33
4	4	3	4,54	5,59	7,14	5	5	3	4,54	5,70	7,57
4	4	4	4,65	5,69	7,65	5	5	4	4,52	5,66	7,82
5	2	2	4,37	5,16	6,53	5	5	5	4,56	5,78	7,98

Источник. [56]. Воспроизведено с разрешения издателя.

Мы применили в иллюстративных целях к данным примера 12.18.1 второй критерий. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [16, с. 191—195], [56], [105, с. 139—142].

12.20. НЕСКОЛЬКО НЕЗАВИСИМЫХ ВЫБОРОК. КРИТЕРИЙ МЕДИАНЫ

Обычная форма записи данных. См. параграф 12.18.

Статистическая модель. Все наблюдения независимы. Выборки извлекаются из k совокупностей, каждая выборка — из одной совокупности.

Гипотезы.

H_0: все k совокупностей имеют одинаковое распределение;

H_1: нулевая гипотеза неверна.

Критическая область. Верхняя 5 %-ная область распределения χ^2_{k-1}.

Вычисление значения критериальной статистики. 1. Объединим все k выборок в единую большую выборку.

2. Найдем *медиану* этой объединенной выборки (см. параграф 8.5).

3. Для каждой выборки подсчитаем число значений, превосходящих общую медиану. Запишем эти числа в таблицу типа табл. 12.20.1.

4. Вычислим ожидаемое значение наблюдений для каждой из $2k$ клеток таблицы. Для этого необходимо умножить элемент строки «всего», соответствующий данной клетке, на элемент столбца «всего», соответствующий ей, и разделить получившееся произведение на полное число наблюдений n.

5. Вычислим значение критериальной статистики по формуле:

$$\chi^2 = \sum_{\substack{\text{все } 2k \\ \text{клеток}}} \frac{(\text{наблюдаемое значение} - \text{ожидаемое значение})^2}{\text{ожидаемое значение}}. \qquad (12.20.1)$$

Комментарии. 1. При условии, что данные упорядочены в форме табл. 12.20.1, данный критерий есть критерий равенства биномиальных вероятностей в нескольких выборках (см. параграф 12.8).

Таблица 12.20.1

	Выборка 1	Выборка 2	...	Выборка k	Всего
Число наблюдений, которые больше медианы	L_1	L_2	...	L_k	$\sum_j L_j$
Число наблюдений, которые меньше медианы	$n_1 - L_1$	$n_2 - L_2$	...	$n_k - L_k$	$n - \sum_j L_j$
Всего	n_1	n_2	...	n_k	n

2. Иногда значения двух или более наблюдений могут быть равны медиане. Тогда их надо поделить поровну между группами, значение наблюдений в которых «выше медианы» и «ниже медианы». Если число таких наблюдений нечетное, они должны быть поделены поровну, а последнее наблюдение надо исключить из анализа.

3. Следует также обратить внимание на критерии, рассмотренные в параграфах 12.18 и 12.19.

4. Если гипотеза H_0 отклоняется и мы хотели бы заключить, что средние распределений различаются, то нужно *предположить,* что изучаемые совокупности одинаковы по всем параметрам, кроме средних.

Пример 12.20.1. Применим метод этого параграфа к данным о длине червей (см. табл. 12.18.2) и проверим нулевую гипотезу, предполагающую, что три распределения длины червей одинаковы.

Для 5 %-ного уровня значимости возьмем верхнюю 5 %-ную область распределения $x|>$ что соответствует значению критерия, превышающему 5,99. Общая медиана равна 9,2, так что в группе 1 только одно из шести наблюдений больше нее, в группе 2 только

одно из семи наблюдений меньше медианы, а в группе *3* два наблюдения из пяти больше медианы. Эти результаты сведены в табл. 12.20.2. В ней также отражены ожидаемые значения, вычисленные по данным строк и столбцов «всего» и общему числу наблюдений (например, 3 = (6X9)/18).

Таблица 12.20.2. **Число червей, длина которых больше и меньше медианы. В скобках показаны ожидаемые значения**

	Выборка 1	Выборка 2	Выборка 3	Всего
Число червей, длина которых больше медианы	1 (3)	6 (3,5)	2 (2,5)	9
Число червей, длина которых меньше медианы	5 (3)	1 (3,5)	3 (2,5)	9
Всего	6	7	5	18

Ожидаемые значения в отдельных клетках достаточно малы. Тем не менее мы вычисляем значение χ^2 обычным путем:

$$\chi^2 = (1 - 3)2/3 + (5 - 3)^2/3 + (6 - 3{,}5)^2/3{,}5 + \ldots$$
$$\ldots + (3 - 2{,}5)^2/2{,}5 = 8{,}4.$$

Эта величина значима для 5 %-ного уровня, поэтому мы отклоняем нулевую гипотезу и приходим к заключению, что наши три распределения не идентичны. Если можно *предположить,* что эти распределения совпадают во всем, кроме средних, то мы сможем утверждать, что не все средние совпадают.

В иллюстративных целях к этим данным мы применили третий критерий. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [7, с. 246—248; русский перевод с. 230—232].

12.21. НЕСКОЛЬКО НЕЗАВИСИМЫХ ВЫБОРОК. МНОЖЕСТВЕННЫЕ СРАВНЕНИЯ ШЕФФЕ

Дисперсионный анализ, рассмотренный в параграфе 12.18, может показать нам, что совокупности, из которых отбираются наблюдения, различаются по своим средним. Однако этот критерий не позволяет узнать, средние каких совокупностей действительно различаются между собой. Можно было бы попытаться произвести серию попарных сравнений совокупностей при помощи t-метода (см. параграф 12.15), но такой подход ошибочен. Если выбрать 5 %-ный уровень значимости, то вероятность отклонить верную гипотезу равна 5 % при каждом сравнении. Когда производится серия сравнений, эта вероятность становится значительно больше 5 %. Для решения возникающей проблемы может служить метод множественного сравнения Шеффе. Этот метод дает возможность

провести попарные сравнения так, что вероятность хотя бы одного неверного заключения в точности равна 5 %.

Обычная форма записи данных. См. параграф 12.18.

Статистическая модель. См. параграф 12.18.

Гипотезы.

$H_0: c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k = 0$, где $\{c_i\}$ — обусловленные постоянные с нулевой суммой;

H_1: нулевая гипотеза неверна.

Критическая область. Выше верхней 5 %-ной точки $F_{k-1,\ n-k}$-распределения.

Вычисление значения критериальной статистики. 1. Вычислим среднее каждой выборки: $x_{.i} = T_{.i}/n_i$.

2. Обозначим остаточный средний квадрат в табл. 12.18.1 через s^2.

3. Вычислим значение критериальной статистики:

$$S = \frac{(c_1\bar{x}_{.1} + c_2x_{.2} + \ldots + c_k\bar{x}_{.k})^2}{(k-1)\,s^2\left(c_1^2/n_1 + \ldots + c_k^2/n_k\right)}. \qquad (12.21.1)$$

Комментарии. 1. Этот метод обычно применяется для серии сравнений типа

$$H_0: \mu_1 - \mu_2 = 0;$$
$$H_1: \mu_1 - \mu_2 \neq 0.$$

2. Существуют другие методы множественного сравнения, например метод Тьюки и метод Стьюдента—Ньюмена—Кейлса (см. [105, с. 151—161]).

3. При проведении серии попарных сравнений можно столкнуться с противоречием *. Во избежание этого явления сравнения необходимо проводить в определенном порядке, а именно следует сравнить группу с наибольшим средним с группой, обладающей наименьшим средним, затем с группой с наименьшим средним среди остальных групп и т. д. Как только при сравнении обнаружится незначимая величина критериальной статистики или не останется группы с меньшим выборочным средним, следует заменить группу с наибольшим средним на группу со вторым по величине средним и начать процедуру сначала.

Пример 12.21.1. С помощью критериев из параграфов 12.18, 12.19 и 12.20 было выяснено, что средняя длина червя для трех групп не одинакова (см. табл. 12.18.2). Воспользуемся методом Шеффе для определения того, какие средние в действительности различаются.

Критическая область есть верхняя 5 %-ная область $F_{2,15}$-распределения, что соответствует значению критериальной статистики

* Например, две величины, порознь равные третьей, окажутся не равны между собой.— *Примеч. пер.*

3,68. Выборочные средние равны 8,60; 10,87 и 9,32 и из табл. 12.18.3 находим, что средний остаточный квадрат s^2 равен 1,395.

Согласно п. 3 комментариев первое сравнение должно быть произведено между выборкой 2 и выборкой 1. Положим $c_1 = -1$, $c_2 = 1$, $cз = 0$ и вычислим значения критериальной статистики:

$$S \quad \frac{(10{,}87 - 8{,}60)^2}{2 \text{ X } 1.395 \text{ X } \left(\frac{1}{6} + \frac{1}{7}\right)} = 5{,}97.$$

Это значение лежит в критической области, поэтому мы отклоняем нулевую гипотезу, предполагающую, что $\mu_1 = \mu_2$, и заключаем, что $\mu_1 \neq \mu_2$.

Согласно п. 3 комментариев следующее сравнение делается между выборками 2 и 3. Мы полагаем, что $c_1 = 0$, $c_2 = 1$ и $Cз = -1$;

$$S \quad \frac{(10{,}87 - 9{,}32)^2}{2\text{X}1.395\text{X}\left(\frac{1}{7} + \frac{1}{6}\right)} = 2{,}51.$$

Эта величина не является значимой, поэтому мы принимаем нулевую гипотезу $H2 = \mu_3$ и заключаем, что $HI \neq \mu_2 = \mu_3$.

Литература: *[3,* с. 176—178], [16, с. 207—212], [39, с. 54—59], [88, с. 68—82; русский перевод с. 72—86], [91, с. 251—253; русский перевод с. 310—313], [105, с. 151—161].

12.22. ПАРНЫЕ НАБЛЮДЕНИЯ. ПАРНЫЙ *t*-КРИТЕРИЙ

Обычная форма записи данных. Часто случается, что значения наблюдений объединены в пары, например:

	Вес до диеты	Вес после диеты
Пациент 1	x_{11}	x_{12}
Пациент 2	x_{21}	x_{22}
.	.	.
.	.	.
.	.	.
Пациент n	x_{n1}	x_{n2}

Статистическая модель. Вес пациента после перехода на диету равен его весу до диеты плюс эффект от ее воздействия. Эффект диеты есть нормально распределенная случайная величина со средним и и дисперсией σ^2.

Гипотезы.

а) *Равенство*
H_0: и = Ио>
$H_1 : y \neq \mu_0$.

б) *Неравенство*
$H_0 : \mu \leqslant \mu_0$;
$H_1 : И > \mu_0$.

в) *Неравенство*
$H_0 : (1 \geqslant \mu_0$;
$H_1 : и < \mu_0$.

Критическая область.

а) *Равенство.* Верхняя 2,5 %-ная область и нижняя 2,5 %-ная область *t*-распределения с $(n - 1)$ степенями свободы.

б) *Неравенство.* Верхняя 5 %-ная область *t*-распределения с $(n - 1)$ степенями свободы.

в) *Неравенство.* Нижняя 5 %-ная область *t*-распределения с $(n - 1)$ степенями свободы.

Вычисление значения критериальной статистики. Вычислим n разностей: $d_1 = x_{12} - x\backslash\backslash$, $d_2 = x_{22} - x_{21}$, ... Вычислим среднюю разность: $d = (d\backslash + \ldots + d_n)/n$.

s_d^2 — выборочная дисперсия для выборки разностей $\{d_i\}$ (формула (8.4.3)).

Положим μ = Но в формуле

$$t = n^{\frac{1}{2}} (\bar{d} - \mu)/s_d. \qquad (12.22.1)$$

Комментарии. 1. Для проверки нулевой гипотезы, состоящей в том, что диета не дает никакого эффекта, следует применить метод а) со значением $\mu_0 = 0$.

2. Если значения наблюдений не объединены в пары, необходимо использовать *t*-критерий из параграфа 12.15.

3. Если статистическая модель, на которой основывается критерий, не соответствует данным наблюдений, то можно применить непараметрические критерии из параграфов 12.23 и 12.24. Заметим, однако, что критерий нечувствителен к умеренным отклонениям от нормальности.

4. Применение критерия а) для проверки равенства эквивалентно дисперсионному анализу по двум признакам для двух столбцов чисел (см. параграф 12.25), если $\mu_0 = 0$.

5. При проверке равенства в силу симметрии *t*-распределения достаточно сравнить $|t|$ с верхней 2,5 %-ной точкой *t*-распределения.

Пример 12.22.1. Два сорта пшеницы сравнивают по урожайности. Сорт *A* — обычная разновидность, а сорт *B* — новый гибрид. Для этого выбирают 10 двухакровых участков и каждый из них делят на два одноакровых. Условия роста и созревания одинаковы на каждом отдельном участке. На каждом двухакровом участке случайным образом выбирают один из одноакровых и засевают его пшеницей сорта *A;* оставшиеся 10 одноакровых участков засевают сортом *B*. Результаты сбора урожая приведены в табл. 12.22.1. Есть ли подтверждение того, что урожайность сорта *B* выше урожайности сорта Л?

Мы должны проверить:

$$H_0 : \mu \leqslant 0;$$
$$H_1 : \mu > 0.$$

5 %-ная критическая область соответствует верхней 5 %-ной области t_9-распределения, что превышает значение критерия 1,833. Десять разностей представлены в табл. 12.22.1. Средняя разность *3*

Таблица 12.22.1. Урожайность пшеницы, бушелей

Участок	Сорт *A*	Сорт *B*	Разность (B − *A*)
1	36,9	36,8	−0,1
2	35,2	37,1	1,9
3	31,2	31,4	0,2
4	34,1	34,1	0,0
5	36,1	35,9	-0,2
6	34,1	35,2	1,1
7	37,2	37,9	0,7
8	36,8	37,2	0,4
9	29,6	30,2	0,6
10	35,4	36,5	1,1

равна 0,57, а выборочная дисперсия разностей s_d^2 равна 0,4312. Значение критерия следующее:

$$t = (10)^{1/2}\ (0{,}57 - 0)/(0{,}4312)^{1/2} = 2{,}7.$$

Эта величина значима, таким образом мы отклоняем нулевую гипотезу и заключаем, что урожайность сорта *B* выше урожайности сорта *A*.

Литература: [3, с. 134—145], [16, с. 167—169], [39, с. 24—26], [91, с. 52—53, 77—80; русский перевод с. 60—62, 86—88], [105, с. 121—124].

12.23. ДВЕ СВЯЗАННЫЕ ВЫБОРКИ. КРИТЕРИЙ ЗНАКОВ

Обычная форма записи данных. См. параграф 12.22. Определим разности:

$$d_1 = x_{12} - x_{11},\ d_2 = x_{22}\ \ x_{21},\ \cdots\ d_n = x_{n2}\ \ x_{n1}.$$

Статистическая модель. Разность d_i является случайно выбранным наблюдением из совокупности *Pi*. Совокупности $\{P_i\}$ ($i = 1, \ldots, n$) имеют одну и ту же медиану.

Гипотезы.

H_0: общая медиана равна нулю;

H_1: общая медиана не равна нулю.

Критическая область.

а) *Малые выборки* ($n \leqslant 20$). Верхняя 2,5 %-ная и нижняя 2,5 %-ная области биномиального распределения с параметрами n и $p = 0{,}5$ (см. табл. 12.23.1).

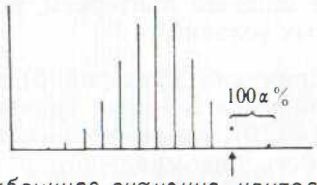
б) *Большие выборки.* Верхняя 2,5 %-ная и нижняя 2,5 %-ная области стандартного нормального распределения.

Вычисление значения критериальной статистики.

а) *Малые выборки.*

Запишем знак разности для каждой пары. Подсчитаем число положительных знаков и обозначим его через *N*. Воспользуемся *N* в качестве критерия.

Таблица 12.23.1. Верхние* 100α %-ные точки биномиального распределения с параметрами n и 0,5 (критическая область включает табличное значение критерия)



n	a = 0,005	0,01	0,025	0,05
5	–	–	–	5
6	–	–	6	6
1	–	7	7	7
8	8	8	8	7
9	9	9	8	8
10	10	10	9	9
11	11	10	10	9
12	11	11	10	10
13	12	12	11	10
14	13	12	12	И
15	13	13	12	12
16	14	14	13	12
17	15	14	13	13
18	15	15	14	13
19	16	15	15	14
20	17	16	15	15

* Нижняя ЮОа %-ная точка получается вычитанием элемента таблицы из биномиального параметра *n*.

б) Большие выборки. Вычислим *N как* в п. а) и положим

$$T = (2N - n)/\sqrt{n}. \qquad (12.23.1)$$

Комментарии. 1. В случае равенства значений в паре, когда разность равна нулю, следует прибавить к *N* величину 0,5.

2. С помощью данного метода могут быть проверены и другие гипотезы. Например:

а) H_0: общая медиана $= \delta$ (δ — обусловленная постоянная);
H_1: общая медиана $\neq \delta$.

Вычтем δ из каждой разности и будем действовать, как раньше.

б) H_0: общая медиана $\leqslant \delta$;
H_1: общая медиана $> \delta$.

Вычтем δ из каждой разности и воспользуемся верхней 5 %-ной областью для биномиального или нормального распределения.

3. Если к некоторым данным применима статистическая модель из параграфа 12.22, то следует применять более мощный t-критерий. Непараметрический критерий Уилкоксона из параграфа 12.24 является также более мощным критерием, но он предполагает выполнение более сильных условий.

Пример 12.23.1. Применим критерий б) из п. 2 комментариев к примеру 12.22.1, принимая 5 %-ный уровень значимости.

Выборка мала ($n=10$), поэтому критическая область есть верхняя 5 %-ная область биномиального распределения с $n=10$ и $p=0{,}5$. Согласно табл. 12.23.1 верхняя 5 %-ная область этого распределения включает $N=9$ и $N=10$. Наблюдаемое значение $N-7{,}5$ не лежит в ней, поэтому мы принимаем нулевую гипотезу о том, что урожайность сорта B не выше урожайности сорта A.

Это заключение противоречит полученному с помощью t-критерия результату и подчеркивает то обстоятельство, что критерий знаков слабее t-критерия.

Данные из примера 12.22.1 исследованы с помощью второго критерия в иллюстративных целях. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [3, с. 229], [7, с. 242—246, русский перевод с. 226-230], [16, с. 153-156], [81, с. 119-131], [86, с. 316], [105, с. 290—291].

12.24. КРИТЕРИЙ УИЛКОКСОНА ДЛЯ ПАРНЫХ ВЫБОРОЧНЫХ НАБЛЮДЕНИЙ

Вместо парного t-критерия из параграфа 12.22 часто используется критерий, рассмотренный в данном параграфе. Это делается в случаях, когда не выполнены предположения, на которых основывается t-критерий (например, выборки получены не из нормальных совокупностей).

Описание критерия будет дано на примере анализа веса пациентов до и после применения определенной диеты. Рассматриваемая нулевая гипотеза состоит в следующем: общее распределение разностей симметрично относительно нуля. Если эта гипотеза отклоняется, то вывод, что средние веса до применения диеты и после нее не равны, можно сделать лишь тогда, когда выполняется предположение о том, что либо распределения разностей имеют один и тот же вид, либо оба они симметричны. На практике, однако, допустимы умеренные отклонения от выполнения этих требований, так как критерий не слишком чувствителен к ним.

Обычная форма **записи данных.** См. параграф 12.22. Число пар обозначается через n. Символом d_i обозначены разности в i-й паре.

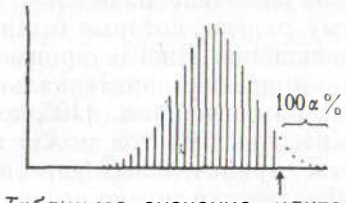
Статистическая модель. Предполагается, что разности независимы и одинаково распределены.

Гипотезы.

H_0: распределение разностей симметрично относительно нуля;

H_1: нулевая гипотеза неверна.

Таблица 12.24.1. **Верхние* 100α %-ные точки распределения Уилкоксона для парных выборок (критическая область включает табличное значение критерия)**



n	a = 0,05	.0,025	0,01	0,005
1	25	26	28	—
8	31	33	35	36
9	37	40	42	44
10	45	47	50	52
И	53	56	59	61
12	61	65	69	71
13	70	74	79	82
14	80	84	90	93
15	90	95	101	105
16	101	107	113	117
17	112	119	126	130
18	124	131	139	144
19	137	144	153	158
20	150	158	167	173

* Нижняя 100α %-ная точка получается вычитанием верхней ЮОа %-ной точки из $0{,}5n(n+1)$.

Источник. [63]. Воспроизведено с разрешения издателя.

Критическая область.

а) *Малые выборки.* Верхняя и нижняя 2,5 %-ные области распределения рангов Уилкоксона (эти ранги могут принимать и отрицательные значения), см. табл. 12.24.1.

б) *Большие выборки.* Верхняя и нижняя 2,5 %-ные области стандартного нормального распределения.

Вычисление значения критериальной статистики.

а) *Малые выборки.* Ранжируем абсолютные величины разностей в возрастающем порядке. Припишем им соответствующие ранги от 1 до я. Каждому значению ранга припишем знак его разности. Вычислим сумму значений положительных рангов и обозначим ее через N. Проверим, принадлежит ли N критической области.

б) *Большие выборки.* Вычислим N как в п. а) и положим

$$T = \{N - n(n+1)/4\}/\{я(л+1)(2л+0/24\}^{\frac{1}{2}}. \qquad (12.24.1)$$

Комментарии. 1. Нулевые разности игнорируются (число n следует при этом соответствующим образом уменьшить).

2. Равным по абсолютной величине разностям следует приписать ранги, равные среднему рангов, которые были бы им приписаны при условии их несовпадения. Иногда производится небольшое уточнение процедуры вычисления критериальной статистики с целью учета совпадающих разностей (см. [105, с. 126]).

3. Вместо суммы положительных рангов может использоваться абсолютная величина суммы отрицательных рангов, при этом получают те же результаты. Фактически можно вычислять обе суммы и сравнивать большую из них с верхним 2,5 %-ным критическим значением.

4. Могут проверяться другие гипотезы. Допустим, что пары наблюдений получены из совокупностей, распределения которых либо симметричны, либо совпадают по всем параметрам, кроме средних. Можно было бы проверить следующие гипотезы:

а) H_0: средняя прибавка в весе = б (δ — обусловленная константа);

H_1: средняя прибавка в весе $\neq \delta$.

Из каждой разности следует вычесть б и действовать, как указано выше.

б) H_0: средняя прибавка в весе $\leqslant \delta$;

H_1: средняя прибавка в весе $> \delta$.

Из каждой разности следует вычесть δ, вычислить значение критериальной статистики и использовать верхнюю 5 %-ную критическую область.

5. Если к некоторым данным применима статистическая модель из параграфа 12.22, то следует пользоваться более мощным t-критерием.

Пример 12.24.1. Применим критерий б) из п. 4 комментариев к примеру, приведенному в параграфе 12.22. Примем 5 %-ный уровень значимости. Положим $\delta = 0$.

Объем выборки сравнительно мал ($n = 10$), при этом одна из разностей равна нулю. Следовательно, необходимо использовать табл. 12.24.1 для $n = 9$. Верхняя 5 %-ная область включает значения $N \geqslant 37$. Ранги с их знаками представлены в табл. 12.24.2.

Таблица 12.24.2. Ранги с приписанными им знаками, соответствующие табл. 12.22.1

Участок	Разность	Ранг со знаком	Участок	Разность	Ранг со знаком
1	−0,1	-1	6	1,1	7,5
2	1,9	9	7	0,7	6
3	0,2	2,5	8	0,4	4
4	0,0	–	9	0,6	5
5	-0,2	-2,5	10	1,1	7,5

Из нее получаем, что $N = 41{,}5$. Это значение лежит в критической области, поэтому мы отклоняем нулевую гипотезу и заключаем, что средняя урожайность сорта *B* выше средней урожайности сорта *A*.

Данные из примера 12.22.1 в иллюстративных целях исследованы с помощью трех различных критериев. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [7, с. 258—260; русский перевод с. 241—244], [16, с. 160—166], [86, с. 318].

12.25. ДИСПЕРСИОННЫЙ АНАЛИЗ ПО ДВУМ ПРИЗНАКАМ ДЛЯ ЗАВИСИМЫХ (ПАРНЫХ) ВЫБОРОК

В данном параграфе рассматривается весьма общий метод, называемый *дисперсионным анализом*. Краткое описание общей процедуры применения этого метода дано в начале параграфа 12.18.

Обычная форма записи данных.

	Столбцы				Всего
	1	2	...	c	
Строка 1	x_{11}	x_{12}	...	x_{1c}	$T_{1.}$
Строка 2	x_{21}	x_{22}	...	x_{2c}	$T'2.$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
Строка r	x_{r1}	x_{r2}	...	x_{rc}	$T_{r.}$
Всего	$T_{.1}$	$T.2$	...	$T.c$	$m..$

Статистическая модель. Значение наблюдения в i-й строке п j-м столбце складывается из следующих элементов: общее среднее значение плюс сумма эффектов i-й строки и j-го столбца плюс некоторая случайная компонента, которая предполагается нормально распределенной. Среднее значение случайной компоненты равно нулю, а дисперсия равна σ^2. Таким образом,

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

где μ — общее среднее наблюдаемых совокупностей; α_i — величина эффекта i-й строки; β_j — величина эффекта j-го столбца; e_{ij} — случайная компонента. Величины $\{e_{ij}\}$ независимы.

Гипотезы.

а) *Равенство эффектов строк*

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_r$;

H_1: не все эффекты равны.

б) *Равенство эффектов столбцов*

$H_0: \beta_1 = \beta_2 = \ldots = \beta_c$;

H_1: не все эффекты равны.

Критическая область.

а) *Эффекты строк.* Верхняя 5 %-ная область F-распределения с $(r-1)$ и $(r-1)(c-1)$ степенями свободы для числителя и знаменателя соответственно.

б) *Эффекты столбцов.* Верхняя 5 %-ная область F-распределения с $(c-1)$ и $(r-1)(c-1)$ степенями свободы для числителя и знаменателя соответственно.

Вычисление значения критериальной статистики. Возведем в квадрат и сложим значения наблюдений; получим $\sum\sum x_{ij}^2$. Вычислим общую сумму квадратов

$$\sum\sum x_{ij}^2 - T_{..}^2/rc.$$

Вычислим сумму квадратов между строками:

$$(T_{1.}^2/c + \ldots + T_{r.}^2/c) - T_{..}^2/rc.$$

Вычислим сумму квадратов между столбцами:

$$(T_{.1}^2/r + \ldots + T_{.c}^2/r) - T_{..}^2/rc.$$

Заполним таблицу дисперсионного анализа [4] (см. табл. 12.25.1).

а) *Эффекты строк.* Значение критериальной статистики равно:

$$F = \frac{\text{средний квадрат между строками}}{\text{остаточный средний квадрат}}. \qquad (12.25.1)$$

б) *Эффекты столбцов.* Значение критериальной статистики равно:

$$F = \frac{\text{средний квадрат между столбцами}}{\text{остаточный средний квадрат}}. \qquad (12.25.2)$$

Таблица 12.25.1. Дисперсионный анализ по двум признакам

Компонента дисперсии (1)	с. к. (2)	с. с. (3)	ср к. (4) = (2)/(3)
Между строками	$(\sum T_{i.}^2/c) - T_{..}^2/rc$	$r-1$	(определяется делением)
Между столбцами	$(\sum T_{.j}^2/r) - T_{..}^2/rc$	$c-1$	
Остаточная	(определяется вычитанием)	$(r-1)(c-1)$	
Полная	$\sum\sum x_{ij}^2 - T_{..}^2/rc$	$rc-1$	—

Комментарии. 1. Проверка равенства из параграфа 12.22 с помощью парного t-критерия представляет собой частный случай дисперсионного анализа по двум признакам (при проверке эффектов столбцов).

2. По-видимому, критерий в разумной степени устойчив, когда распределения совокупностей умеренно отличаются от нормальных или имеется умеренное отклонение от выполнения требования равенства дисперсий. Для проверки величины отклонений совокупностей от нормальности и проверки значимости различия величин дисперсий можно действовать следующим образом. Прибавим к значению каждого наблюдения общее среднее всех совокупностей $x_{..} = T_{..}/rc$ и вычтем сумму средних значений в столбце и строке, содержащих данное наблюдение, $x_{i.} = T_{i.}/c$ и $x_{.j} = T_{.j}/r$. После этого следует проверить гипотезу, состоящую в том, что получившиеся числа являются выборкой из нормальной совокупности с нулевым средним и дисперсией, равной остаточному среднему квадрату.

3. Если рассматриваемые данные не удовлетворяют предположениям статистической модели дисперсионного анализа по двум признакам, то необходимо применить непараметрический критерий из параграфа 12.26.

4. Могут проверяться другие гипотезы. Например:

в) $H_0: \alpha_1 - \alpha_2 = 6;\ \alpha_2 = \alpha_3 = \ldots = \alpha_r;$

H_1: нулевая гипотеза неверна.

Следует воспользоваться методом из п. а) после вычитания шести из каждого элемента в строке 1.

г) H_0:р, - $\beta_2 = 5;\ \beta_2 = \beta_3 = \ldots = \beta_c;$

H_1: нулевая гипотеза неверна.

Следует воспользоваться методом из п. б) после вычитания пяти из каждого элемента в столбце 1.

5. При вычислении сумм квадратов между строками и между столбцами возводятся в квадрат соответствующие элементы строки или столбца. При этом полученная сумма квадратов делится на число элементов этой суммы.

6. Когда критерий дает значимый результат при проверке равенства эффектов строк (столбцов), то можно заключить, что не все эффекты строк (столбцов) равны. Для того чтобы определить, эффекты каких строк (столбцов) действительно различаются, нужно применить метод множественного сравнения (см. параграф 12.27). Не следует сравнивать строки (столбцы) с помощью парного t-критерия или дисперсионного анализа.

7. *Одно пропущенное значение.* В экспериментальных данных, полученных даже в лучших биологических лабораториях, могут встретиться пропуски (например, разбилась колба или животные погибли). Для решения проблем, возникающих из-за пропущенного значения, можно применить метод, изложенный далее. Заметим, однако, что его нельзя применять, если пропуск значения прямое следствие эксперимента (например, животное погибло из-за отравления исследуемой диетой). Пусть пропущено одно значение в таблице дисперсионного анализа, стоящее в i-й строке и j-м столбце. Тогда нужно приписать соответствующему элементу таблицы значение

$$X = (rT'_{i.} + cT'_{.j} - T'_{..})/\{(r-1)(c-1)\}. \qquad (12.25.3)$$

[4] См. сноску 3 на с. 175.

Здесь $T'_{i.}$, $T'_{.j}$ и $T'_{..}$ — соответственно элемент столбца «всего», элемент строки «всего» и сумма элементов столбца «всего» (или, что то же самое, строки «всего») в таблице, где на месте пропущенного значения стоит 0. После такого приписывания проводят обычную процедуру дисперсионного анализа по двум признакам с числом степеней свободы, уменьшенным на единицу. Описанная процедура является приближенной. Существует также точный метод, позволяющий, например, проверить гипотезу о равенстве эффектов строк. При его применении можно действовать следующим образом.

Таблица **12.25.2. Дисперсионный анализ по двум признакам с p пропущенными значениями**

Компонента дисперсии (1)	с. к. (2)	с. с. (3)	ср. к. (4)=(2)/(3)
Между столбцами	(см. п. 7 комментариев)	$c-1$	(определяется делением)
Дополнительное уменьшение, обусловленное строками		$r-1$	
Остаточная		$(r-1)(c-1)-p$	
Полная	(см. п. 7 комментариев)	$rc-1-p$	—

а) Вычислить сумму квадратов между строками и общую сумму квадратов с помощью вычислительной процедуры, выполняемой при проведении дисперсионного анализа по одному признаку (см. параграф 12.18). При этом пропущенное(ые) значение(ия) игнорируется(ются).

б) Вычислить остаточную сумму квадратов с помощью вычислительной процедуры, выполняемой при проведении дисперсионного анализа по двум признакам. При этом пропущенное значение (или пропущенные значения, если их несколько) заменяется на его оценку (или их оценки), полученную, как описано выше.

в) Вычислить дополнительное уменьшение суммы квадратов, связанной со строками, путем соответствующего вычитания.

Табл. 12.25.2 соответствует дисперсионному анализу по двум признакам с p пропущенными значениями. Критическая область дается F-распределением с $(r-1)$ и $[(r-1)(c-1)-1]$ степенями свободы (для одного пропущенного значения). Для проверки гипотезы о различии столбцов следует повторить всю процедуру, поменяв местами строки и столбцы.

8. *Два пропущенных значения.* Для разбора этого случая можно обратиться к следующему итеративному процессу. Заменим пропущенное значение произвольным числом, затем с помощью получившейся таблицы оценим второе пропущенное значение, как описано выше. Далее произведем оценку первого пропущенного значения, используя полученную оценку второго. Эту процедуру

будем проводить итеративно, пока не придем к решению (т. е. полученные оценки величин пропущенных значений перестанут сильно меняться). После получения полной таблицы следует применить метод, рассмотренный в п. 7 комментариев, уменьшив число степеней свободы на два по сравнению с обычным.

Пример 12.25.1. Было определено содержание фосфора (мг/100 г) в каждом из четырех органов у некоторых животных трех пород. Результаты экспериментов представлены в табл. 12.25.3. Требуется проверить гипотезу, состоящую в том, что у данных животных не существует различия в уровне содержания фосфора.

Таблица **12.25.3. Содержание фосфора (мг/100 г) в органах животных**

	Сердце	Легкие	Печень	Почки	Всего
Порода 1	86,7	102,7	204,6	184,6	578,6
Порода 2	88,4	108,1	213,2	183,4	593,1
Порода 3	81,2	99,8	201,1	179,0	561,1
Всего	256,3	310,6	618,9	547,0	1732,8

Критическая область для 5 %-ного уровня значимости есть верхняя 5 %-ная область $F_{2,6}$-распределения. Она определяется неравенством $F > 5{,}14$. Сумма квадратов значений наблюдений в таблице равна 281 628,16. Поэтому общая сумма квадратов следующая:

$$281\,628{,}16 - (1732{,}8)^2/12 = 31\,411{,}840.$$

Сумма квадратов между строками равна:

$$(578{,}6)^2/4 + (593{,}1)^2/4 + (561{,}1)^2/4 - (1732{,}8)^2/12 = 128{,}375.$$

Сумма квадратов между столбцами равна:

$$(256{,}3)^2/3 + (310{,}6)^2/3 + (618{,}9)^2/3 + (547{,}0)^2/3 - (1732{,}8)^2/12 = 31\,253{,}100.$$

Таблица 12.25.4. **Дисперсионный анализ содержания фосфора в органах животных**

Компонента дисперсии	с. к.	с. с.	ср. к.
Строки	128,375	2	64,188
Столбцы	31 253,100	3	10 417,700
Остаточная	30,365	6	5,061
Полная	31 411,840	И	—

Теперь можно заполнить табл. 12.25.4. Она содержит результаты дисперсионного анализа имеющихся данных. Значение критериальной статистики равно:

$$F = 64{,}188/5{,}061 = \mathbf{12{,}7}.$$

Эта величина лежит в критической области; следует отклонить нулевую гипотезу и заключить, что у животных разных пород существует различие в содержании фосфора.

Литература: [3, с. 165—173], [7, с. 467—501], [9, с. 344], [16, с. 195-199], [39, с. 65-84], [45, с. 256—260], [70, с. 374-378], [86, с. 288—300], [88, с. 331—368; русский перевод с. 256—286], [91, с. 291—327; русский перевод с. 275—309], [105, с. 163—174].

12.26. НЕПАРАМЕТРИЧЕСКИЙ ДИСПЕРСИОННЫЙ АНАЛИЗ ФРИДМАНА ПО ДВУМ ПРИЗНАКАМ ДЛЯ ЗАВИСИМЫХ ВЫБОРОК

В случаях когда не выполнены предположения, на которых основан обыкновенный дисперсионный анализ по двум признакам (см. параграф 12.25), часто пользуются непараметрическим критерием, описанным в данном параграфе. Цель его применения (как и обыкновенного дисперсионного анализа) — проверка гипотез об эффектах строк или столбцов. Этот метод следует применять, например, если исследуемые совокупности не нормальны. Он также целесообразен, если данные представлены в виде рангов или порядковых номеров.

Обычная форма записи данных. См. параграф 12.25 (таблица с r строками и c столбцами).

Статистическая модель. Наблюдения независимы. Наблюдение в строке i и столбце j берется из совокупности с функцией распределения $\{F_{ij}(x)\}$ и средним μ_{ij}. В пределах каждой строки i функции $\{F_{ij}(x)\}$ совпадают по всем параметрам, кроме, может быть, средних. Величина среднего значения может зависеть от номера соответствующего столбца. Эффекты столбцов таковы, что если все c средних в каждой строке упорядочены по возрастанию от 1 до c, то порядок одинаков во всех строках. (Таким образом, если $\mu_{13} > \mu_{14}$, то $\mu_{53} > \mu_{54}$.)

Гипотезы.

H_0: эффекты столбцов отсутствуют;

H_1: нулевая гипотеза неверна.

Критическая область.

а) *Малые выборки.* Верхняя 5 %-ная область распределения Фридмана (см. табл. 12.26.1).

б) *Большие выборки.* Верхняя 5 %-ная область распределения χ^2_{c-1}.

Вычисление значения критериальной статистики. Ранжируем элементы в строках от 1 до c; обозначим сумму рангов в j-м столбце через R_j $(j = 1, \ldots, c)$. Вычислим

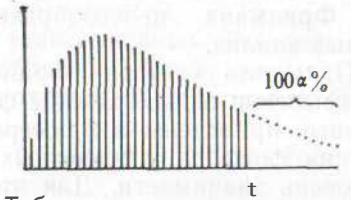
$$S = (R_1^2 + \ldots + R_c^2) - (R_1 + \ldots + R_c)^2/c.$$

Значение критериальной статистики равно:

$$\chi^2 = 12S/\{rc(c+1)\}. \qquad (12.26.1)$$

Комментарии. 1. Если два или более наблюдения имеют одинаковые значения, они называются *совпадающими.* Каждому из

Таблица **12.26.1. Верхние 100α %-ные точки распределения Фридмана (критическая область включает табличное значение критерия)**



Табличное значение критерия

c	r	$\alpha = 0{,}05$	0,02	0,01	0,005	0,001
3	3	6,000	—	—	—	—
3	4	6,500	8,000	8,000	8,000	—
3	5	6,400	6,400	8,400	10,000	10,000
3	6	7,000	8,333	9,000	10,333	12,000
3	7	7,143	8,000	8,857	10,286	12,286
3	8	6,250	7,750	9,000	9,750	12,250
3	9	6,222	8,000	8,667	10,667	12,667
4	2	6,000	—	—	—	—
4	3	7,400	8,200	9,000	9,000	—
4	4	7,800	8,400	9,600	10,200	11,100

Источник, Friedman (1937, J. American Statistical Association, 32, с. 675—701). Воспроизведено с разрешения издателя.

таких наблюдений следует приписать ранг, равный среднему рангов, которые были бы им приписаны при отсутствии совпадения. Иногда производится небольшое уточнение процедуры вычисления критерия для учета совпадающих значений (см. [105, с. 176]).

2. В случае когда отклоняется нулевая гипотеза, нельзя указать, какой столбец оказывает ненулевое влияние. Недопустимо применять данный критерий (или критерии из параграфов 12.23 и 12.24) для сравнения всевозможных пар столбцов. Следует прибегнуть к методу множественного сравнения (см. [105, с. 176—177].

3. Статистическая модель дисперсионного анализа из параграфа 12.25 требует, чтобы эффекты столбцов были одинаковы во всех строках. Статистическая модель критерия Фридмана предполагает выполнение более слабого условия, заключающегося в том, что эффекты столбцов для разных строк имеют одинаковое направление возрастания.

I. Статистическая модель Фридмана предполагает, что функции распределения в каждой строке идентичны во всем, за возможным исключением средних значений. На практике допустимо пренебрегать умеренными отклонениями от этого требования.

5. Для проверки гипотез об эффектах строк следует поменять местами строки и столбцы.

6. Если пригодна статистическая модель из параграфа 12.25, то вместо критерия Фридмана целесообразно применить более мощный дисперсионный анализ.

Пример 12.26.1. Применим критерий Фридмана к экспериментальным данным, описывающим содержание фосфора у животных трех пород. Эти данные приведены в примере 12.25.1. Проверим различие в содержании фосфора у животных разных пород, используя 5 %-ный уровень значимости. Для этого необходимо поменять местами столбцы и строки таблицы.

Согласно табл. 12.26.1 критическая область соответствует значению S, которое больше или равно 6,500. Значения рангов приведены в табл. 12.26.2; по этим данным получим:

$$S = (9^2 + 11^2 + 4^2) - (9 + 11 + 4)^2/3 = 26{,}0.$$

Значение критериальной статистики равно:

$$\chi^2 = (12 \times 26{,}0)/(4 \times 3 \times 4) = 6{,}5.$$

Эта величина в точности значима. Поэтому мы отклоняем нулевую гипотезу и заключаем, что среднее содержание фосфора у животных трех пород не одинаково.

Таблица **12.26.2. Ранжирование данных из параграфа 12.25 о содержании фосфора в органах животных трех пород**

	Порода 1	Порода 2	Порода 3
Сердце	2	3	1
Легкие	2	3	1
Печень	2	3	1
Почки	3	2	1
Всего	9	11	4

К одним и тем же данным в иллюстративных целях были применены два различных критерия. Обращаем внимание читателя на предупреждение, сделанное в параграфе 12.2.

Литература: [3, с. 233], [7, с. 260—262; русский перевод с. 244—246], [16, с. 200—203], [105, с. 175—176].

12.27. ЗАВИСИМЫЕ (ПАРНЫЕ) ВЫБОРКИ. МНОЖЕСТВЕННЫЕ СРАВНЕНИЯ ПО ШЕФФЕ

Дисперсионный анализ по двум признакам (см. параграф 12.25) позволяет обнаружить существование эффектов столбцов в таблице дисперсионного анализа. Однако он не дает возможности точно указать столбцы, которые обладают нулевыми эффектами. Можно было бы попытаться провести серию попарных сравнений с помощью t-критерия (см. параграф 12.22), но такой подход ошибочен. Причина этого объяснена в начале параграфа 12.21. Проведение попарных сравнений возможно методом множественного сравнения Шеффе.

Обычная форма записи данных. См. параграф 12.25 (таблица с r строками и c столбцами).

Статистическая модель. См. параграф 12.25.

Гипотезы.

H_0: $c_1\alpha_1 + c_2\alpha_2 + \ldots + c_r\alpha_r = 0$, где $\{c_i\}$ — определенные константы с нулевой суммой;

H_1: нулевая гипотеза неверна.

Критическая область. Выше, чем верхняя 5 %-ная точка F-распределения с $r - 1$ и $(r - 1)(c - 1)$ степенями свободы.

Вычисление критериальной статистики. 1. Вычислим среднее для каждой строки: $x_{i.} = T_{i.}/c$.

2. Обозначим остаточный средний квадрат в табл. 12.25.1 через s^2.

3. ВЫЧИСЛИМ $$S = \frac{(c_1\bar{x}_{1.} + c_2\bar{x}_{2.} + \ldots + c_r x_{r.})^2}{(r-1)\,s^2\,(c_1^2/c + \ldots + c_r^2/c)}. \qquad (12.27.1)$$

Комментарии. 1. Этот метод обычно применяется для серии сравнений типа

$$H_0: \alpha_1 - \alpha_2 = 0;$$
$$H_1: \alpha_1 - \alpha_2 \neq 0.$$

2. Для того чтобы выполнить попарные сравнения между столбцами, необходимо поменять местами параметры r и c и использовать в вычислениях вместо средних значений для строк средние значения для столбцов.

3. Существуют другие методы множественного сравнения, например метод Тьюки и метод Стьюдента—Ньюмена—Кейлса (см. [39], [105, с. 174]).

4. Следует обратить внимание на п. 3 комментариев в параграфе 12.21.

Пример 12.27.1. Проверки гипотез, выполненные в параграфах 12.25 и 12.26, обнаруживают значительную разницу в содержании фосфора у животных трех пород (см. табл. 12.25.3). С помощью метода Шеффе выясним, для каких пород содержание фосфора действительно различно.

Критическая область совпадает с верхней 5 %-ной областью $F_{2,6}$-распределения. Эта область определяется неравенством $S > 5{,}14$. Средние значения в строках равны соответственно 144,65; 148,275 и 140,275; остаточный средний квадрат $s^2 = 5{,}061$.

Согласно п. 4 комментариев первое сравнение должно быть сделано между строкой 2 и строкой 3. Полагаем $c_1 = 0$, $c_2 = 1$, $c_3 = -1$ и вычисляем значение критерия:

$$S = \frac{(148{,}275 - 140{,}275)2}{2 \times 5{,}061 \times \left(\frac{1}{4} + \frac{1}{4}\right)} = 12{,}6.$$

Это значение лежит в критической области, и мы заключаем, что «2 ≠ α3.

Следующее сравнение производится между строками 2 и 1. Полагаем $c_1 = -1$, $c_2 = 1$, $cз = 0$ и вычисляем:

$$\delta = \frac{(148{,}275 - 144{,}65)^2}{2 \times 5{,}061 \times \left(\frac{1}{4} + \frac{1}{4}\right)} = 2{,}6.$$

Эта величина не является значимой. Следовательно, можно прийти к выводу, что $\alpha_1 = \alpha_2 \neq \alpha_3$.

Литература: [3, с. 176—178], [16, с. 207—212], [39, с. 79], [88, с. 68—82; русский перевод с. 72—86], [105, с. 151—161].

12.28. КРИТЕРИЙ ДЛЯ ПРОВЕРКИ ВЕЛИЧИНЫ ЕДИНСТВЕННОЙ ДИСПЕРСИИ

Обычная форма записи данных. $x_1, x_2, xз, \ldots, x_n$ — выборка объема п.

Статистическая модель. Наблюдения независимы и выбираются из нормальной совокупности со средним μ и дисперсией σ^2.

Гипотезы.

а) *Равенство*	б) *Неравенство*	в) *Неравенство*
$H_0: \sigma^2 = \sigma_0^2$;	$H_0: \sigma^2 \leqslant \sigma_0^2$;	$H_0: \sigma^2 \geqslant \sigma_0^2$;
$H_1: \sigma^2 \neq \sigma_0^2$.	$H_1: \sigma^2 > \sigma_0^2$.	$H_1: \sigma^2 < \sigma_0^2$.

Критическая область.

а) *Равенство.* Верхняя 2,5 %-ная область распределения χ^2_{n-1}, нижняя 2,5 %-ная область того же распределения.

б) *Неравенство.* Верхняя 5 %-ная область распределения χ^2_{n-1}.

в) *Неравенство.* Нижняя 5 %-ная область распределения χ^2_{n-1}.

Вычисление значения критериальной статистики.

s^2 — выборочная дисперсия совокупности (формула (8.4.3)).

Положим $\sigma^2 = \sigma_0^2$ в формуле

$$\chi^2 = (n - 1)\, s^2/\sigma^2. \qquad (12.28.1)$$

Комментарии. 1. Величина σ_0^2 — обусловленное число (например, 3,67).

2. Критерий не является устойчивым при отклонениях от нормальности. Критерий проверки совокупности на нормальность рассмотрен в параграфе 12.35.

Пример 12.28.1. Проводится статистический анализ характеристик одиннадцати ручных гранат. Требуется проверить утверждение их изготовителя о том, что среднее время срабатывания взрывателя равно 4,01 с, а его стандартное отклонение равно 0,07 с. Получены следующие результаты испытаний: 4,21; 4,03; 3,99; 4,05; 3,89; 3,98; 4,01; 3,92; 4,23; 3,85 и 4,20. Какой вывод можно сделать?

Среднее время срабатывания взрывателя было проверено к примере 12.13.1, где были вычислены $\bar{x} = 4{,}033$ и $s^2 = 0{,}0170$. Теперь желательно проверить гипотезы:

$$H_0: \sigma^2 \leqslant 0{,}0049;$$

$$H_1: \sigma^2 > 0{,}0049.$$

5 %-ная критическая область есть верхняя 5 %-ная область распределения χ^2_{10}, что соответствует критическому значению 18,31. Значение критериальной статистики следующее:

$$\chi^2 = \frac{10 \times 0.0170}{0{,}0049} = 34{,}7.$$

Эта величина (в высокой степени) значима, поэтому мы отвергаем утверждение изготовителя.

К данным примера 12.13.1 в иллюстративных целях здесь был применен второй критерий. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [3, с. 108], [9, с. 326], [39, с. 16—17], [70, с. 307], [86, с. 200—209], [91, с. 59—62; русский перевод с. 58—60], [105, с. 98].

12.29. РАВЕНСТВО (НЕРАВЕНСТВО) ДВУХ ДИСПЕРСИЙ

Обычная форма записи данных. Имеются две выборки, как в параграфе 12.15. Объем первой выборки равен n_1; ее выборочная дисперсия равна s_1^2. Вторая выборка имеет объем n_2, ее выборочная дисперсия равна s_2^2.

Статистическая модель. Первая выборка производится из нормальной совокупности со средним μ_1 и дисперсией σ_1^2. Вторая выборка — из нормальной совокупности со средним μ_2 и дисперсией σ_2^2. Все наблюдения независимы.

Гипотезы.

а) *Равенство*	б) *Неравенство*
$H_0: \sigma_1^2 = \sigma_2^2$;	$H_0: \sigma_1^2 \leqslant \sigma_2^2$;
$H_1: \sigma_1^2 \neq \sigma_2^2$.	$H_1: \sigma_1^2 > \sigma_2^2$.

Критическая область.

л) *Равенство.* Верхняя 2,5 %-ная область $F_{p,q}$-распределения (определение p и q см. далее).

б) *Неравенство.* Верхняя 5 %-ная. область $F_{n_1-1,\, n_2-1}$-распределения.

Вычисление значения критериальной статистики. Следует положить σ^2 равным σ_2^2 в формуле

$$F = \frac{s_1^2}{\sigma_1^2} \Big/ \frac{s_2^2}{\sigma_2^2}. \qquad (12.29.1)$$

Определение p и q в случае а). Если значение критерия больше единицы, возьмем $p = n_1 - 1$ и $q = n_2 - 1$, если оно меньше единицы, то возьмем обратную величину, полагая $p = n_2 - 1$ и $q = n_1 - 1$.

Комментарии. 1. Тот факт, что для проверки равенства надо иногда брать в качестве критерия обратную величину F, вытекает из определения F-распределения (см. параграф 9.4).

2. Могут проверяться другие гипотезы. Например:

в) $H_0: \sigma_1^2 = k\sigma_2^2$, где k — константа;
$H_1: \sigma_1^2 \neq k\sigma_2^2$.

Применяем метод а) и полагаем $\sigma_1^2 = k\sigma_2^2$ в формуле (12.29.1).

г) $H_0: \sigma_1^2 \leqslant k\sigma_2^2$;
$H_1: \sigma_1^2 > k\sigma_2^2$.

Применяем метод б) и полагаем $\sigma_1^2 = k\sigma_2^2$ в формуле (12.29.1).

Пример 12.29.1. В параграфах 12.15 и 12.16 сравнивалась урожайность двух сортов пшеницы. Можно ли утверждать равенство дисперсий для урожайности сорта A и урожайности сорта B?

Следует проверить

$$H_0: \sigma_A^2 = \sigma_B^2;$$
$$H_1: \sigma_A^2 \neq \sigma_B^2.$$

Из содержания параграфа 12.15 можно заключить, что $n_A = n_B = 25$; $s^2 = 5{,}9$ и $s_B^2 = 11{,}2$. Критическая область для 5 %-ного уровня значимости — 2,5 %-ная верхняя область $F_{24,24}$-распределения. Она определяется неравенством $F > 1{,}98$. Значение критерия равно:

$$F = 11{,}2/5{,}9 = 1{,}90.$$

Это значение величины F не лежит в критической области, поэтому можно заключить, что дисперсии равны.

К данным примера 12.15.1 в иллюстративных целях был применен третий критерий. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [3, с. 145—150], [7, с. 282—288; русский перевод с. 256—259], [9, с. 327], [39, с. 18—19], [86, с. 210—215], [91, с. 96; русский перевод с. 105], [105, с. 101—102].

12.30. КРИТЕРИЙ БАРТЛЕТТА ДЛЯ ПРОВЕРКИ РАВЕНСТВА НЕСКОЛЬКИХ ДИСПЕРСИЙ

Обычная форма записи данных. Заданы k выборок, как в параграфе 12.18. Объем i-й выборки равен n_i. Выборочная дисперсия i-й выборки равна s_i^2. Полагаем $n = n_1 + n_2 + \ldots + n_k$.

Статистическая модель. Выборки производятся из нормальной совокупности. Среднее первой совокупности равно μ_1, ее дисперсия равна σ_1^2. Среднее второй совокупности равно μ_2, ее дисперсия равна σ_2^2 и т. д. Все наблюдения независимы.

Гипотезы.

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2;$$

H_1: не все дисперсии равны между собой.

Критическая область. Верхняя 5 %-ная область распределения χ^2_{k-1}.

Вычисление значения критериальной статистики.

Полагаем

$$C = 1 + \{1/(n_1 - 1) + \ldots + 1/(n_k - 1) - 1/(n - k)\}/\{3(k - 1)\};$$
$$S^2 = \{(n_1 - 1)s_1^2 + \ldots + (n_k - 1)s_k^2\}/(n - k).$$
$$\chi^2 = [(n - k)\ln S^2 - \{(n_1 - 1)\ln s_1^2 + \ldots + (n_k - 1)\ln s_k^2\}]/C. \qquad (12.30.1)$$

Комментарии. 1. Этот критерий является приближенным.

2. Если $k = 2$, то следует воспользоваться точным критерием из параграфа 12.29.

3. Чем больше отличаются одна от другой выборочные дисперсии, тем больше значение критериальной статистики (12.30.1). Эта величина мала, если все они приблизительно равны.

4. Могут проверяться другие гипотезы. Например:

$$H_0: \sigma_1^2 = 2\sigma_2^2, \quad \sigma_2^2 = \sigma_3 = \ldots = 01;$$

H_1: нулевая гипотеза неверна.

Следует применять метод Бартлетта с заменой s_i^2 на $s_i^2/2$.

Г). Если использование натуральных логарифмов почему-либо может вызвать трудности, можно прибегнуть в формуле (12.30.1) к десятичным логарифмам. При этом получившееся значение необходимо умножить на 2,3026.

П. Критерий Бартлетта не устойчив при отклонениях от нормальности. Критерий проверки совокупности на нормальность рассмотрен в параграфе 12.35.

Пример 12.30.1. Применим критерий Бартлетта для проверки данных в табл. 12.18.2 на однородность[5] дисперсий. Примем 5 %-ный уровень значимости.

[5] Иногда вместо термина *однородность дисперсии* употребляется термин *гомоскедастичность*, а вместо термина *неоднородность дисперсии* — *гетероскедастичность*.

Критическая область есть верхняя 5%-ная область распределения. Находим, что $s_1^2 = 0{,}7240$; $s_2^2 = 2{,}5657$; $s_3^2 = 0{,}4770$.

$$C = 1 + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{4} - \frac{1}{15}\right)/(3 \times 2) = 1{,}091\,67;$$

$$S^2 = (5 \times 0{,}7240 + 6 \times 2{,}5657 + 4 \times 0{,}4770)/15 = 1{,}3948;$$

$$\chi^2 = \{15 \ln (1{,}3948) - 5 \ln (0{,}7240) - 6 \ln (2{,}5657) - 4 \ln (0{,}4770)\}/1{,}091\,67 = 3{,}585.$$

Эта величина не является значимой, и у нас нет свидетельства неоднородности дисперсий.

К данным табл. **12.18.2** был применен четвертый критерий. Следует обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [5], [7, с. 290—295; русский перевод с. 264—270], [9, с. 328], [39, с. 20—21], [87], [91, с. 285—289; русский перевод с. 271—274], [105, с. 131—133].

12.31. ПРЕОБРАЗОВАНИЕ ФИШЕРА ДЛЯ ПРОВЕРКИ ГИПОТЕЗ О ВЗАИМОЗАВИСИМОСТИ

Обычная форма записи данных. Имеются n точек:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Статистическая модель. Все точки представляют собой случайные выборки из двумерного нормального распределения [6] с коэффициентом корреляции [7] ρ. Выборки независимы.

Гипотезы.

а) *Равенство*	*б) Неравенство*	в) *Неравенство*
$H_0: \rho = \rho_0;$	$\rho \leqslant \rho_0;$	$\rho \geqslant \rho_0;$
$H_1: \rho \neq \rho_0.$	$\rho > \rho_0.$	$\rho < \rho_0.$

Критическая область.

а) *Равенство.* Выше верхней 2,5 %-ной точки и ниже нижней 2,5 %-ной точки стандартного нормального распределения.

б) *Неравенство.* Выше верхней 5 %-ной точки стандартного нормального распределения.

в) *Неравенство.* Ниже нижней 5 %-ной точки стандартного нормального распределения.

---

[6] Двумерное нормальное распределение определено в параграфе 9.6.

[7] Коэффициент корреляции определен в параграфе 8.8.

Вычисление значения критериальной статистики. Вычислим выборочный коэффициент корреляции r (см. параграф 8.8). Положим

$$z = \frac{1}{2} \ln \{(1 + r)/(1 - r)\},$$

$$\zeta_0 = \frac{1}{2} \ln \{(1 + \rho_0)/(1 - \rho_0)\}.$$

Значение критерия равно:

$$T = (n - 3)^{\frac{1}{2}} (2 - \zeta_0). \qquad (12.31.1)$$

Комментарии. 1. Этот критерий приближенный. Точное распределение (зависящее от ρ и от n) является весьма сложным (см. например, [53, с. 387; русский перевод с. 370]).

2. Существует точный критерий проверки нулевой гипотезы, $\rho = 0$, и ее альтернативы, $\rho \neq 0$. Критическая область совпадает с верхней и нижней 2,5 %-ными областями t_{n-2}-распределения; значение критериальной статистики равно:

$$t = (n - 2)^{\frac{1}{2}} r/(1 - r^2)^{\frac{1}{2}}. \qquad \textbf{(12.31.2)}$$

3. 2-преобразование асимптотически нормально и тогда, когда основное распределение не является двумерным нормальным распределением. Однако следует учитывать, что в этих случаях его сходимость к нормальному распределению менее быстрая (это означает, что для достижения определенной точности требуется бо́льшее n). Величина асимметрии предельных распределений незначительно влияет на устойчивость критерия. В то же время эффект отклонения значения эксцесса от эксцесса нормального распределения ($\beta_2 = 3$) может быть значительным. Для получения двумерной совокупности с нормальным предельным распределением обеих компонент могут применяться различные виды преобразования данных.

4. В параграфе 12.34 описан непараметрический метод проверки гипотезы некоррелированности данных (нулевой корреляции).

5. Для получения значений 2-преобразования может служить табл. 12.31.1.

6. В части III нашей книги описаны критерии, связанные с регрессионным анализом.

Пример 12.31.1. В табл. 12.3.2 представлены оценки, полученные пятнадцатью студентами за ответы на два вопроса на экзамене по статистике. Проверим гипотезу о том, что корреляция между оценками за ответы на вопрос 1 и на вопрос 2 равна 0,5.

Таблица 12.31.1. **Значения z-преобразования Фишера**

$$z = \frac{1}{2} \ln \{(1 + r)/(1 - r)\}$$

r	z	r	z	r	z	r	z
0,00	0,000	0,25	0,255	0,50	0,549	0,75	0,973
0,01	0,010	0,26	0,266	0,51	0,563	0,76	0,996
0,02	0,020	0,27	0,277	0,52	0,576	0,77	1,020
0,03	0,030	0,28	0,288	0,53	0,590	0,78	1,045
0,04	0,040	0,29	0,299	0,54	0,604	0,79	1,071
0,05	0,050	0,30	0,310	0,55	0,618	0,80	1,099
0,06	0,060	0,31	0,321	0,56	0,633	0,81	1,127
0,07	0,070	0,32	0,332	0,57	0,648	0,82	1,157
0,08	0,080	0,33	0,343	0,58	0,662	0,83	1,188
0,09	0,090	0,34	0,354	0,59	0,678	0,84	1,221
0,10	0,100	0,35	0,365	0,60	0,693	0,85	1,256
0,11	0,110	0,36	0,377	0,61	0,709	0,86	1,293
0,12	0,121	0,37	0,388	0,62	0,725	0,87	1,333
0,13	0,131	0,38	0,400	0,63	0,741	0,88	1,376
0,14	0,141	0,39	0,412	0,64	0,758	0,89	1,422
0,15	0,151	0,40	0,424	0,65	0,775	0,90	1,472
0,16	0,161	0,41	0,436	0,66	0,793	0,91	1,528
0,17	0,172	0,42	0,448	0,67	0,811	0,92	1,589
0,18	0,182	0,43	0,460	0,68	0,829	0,93	1,658
0,19	0,192	0,44	0,472	0,69	0,848	0,94	1,738
0,20	0,203	0,45	0,485	0,70	0,867	0,95	1,832
0,21	0,213	0,46	0,497	0,71	0,887	0,96	1,946
0,22	0,224	0,47	0,510	0,72	0,908	0,97	2,092
0,23	0,234	0,48	0,523	0,73	0,929	0,98	2,298
0,24	0,245	0,49	0,536	0,74	0,950	0,99	2,647

Источник. Hoel P. G. Elementary statistics. Wiley. New York, 1960. Печатается с разрешения издателя.

Таблица 12.31.2. **Оценки, полученные пятнадцатью студентами за ответы на два экзаменационных вопроса**

Студент	Оценка за вопрос 1	Оценка за вопрос 2	Студент	Оценка за вопрос 1	Оценка за вопрос 2
1	27	15	9	15	10
2	18	5	10	12	7
3	15	10	11	13	8
4	10	9	12	17	14
5	3	2	13	19	13
6	18	10	14	17	11
7	6	8	15	9	6
8	15	9			

Критическая область есть $|T| > 1{,}96$; выборочный коэффициент корреляции равен $r = 0{,}764$. Вычисляем:

$$z = \frac{1}{2} \ln (1{,}764/0{,}236) = 1{,}006;$$

$$\zeta_0 = \frac{1}{2} \ln (1{,}500/0{,}500) = 0{,}549 .$$

Значение критериальной статистики равно:

$$T = (12)^{1/2}(1{,}006 - 0{,}549) = 1{,}583.$$

Эта величина не является значимой, поэтому мы принимаем нулевую гипотезу.

Литература: *[2, с.* 78—90; русский перевод с. 111—127], [3, с. 196—206], [7, с. 397—417], [9, с. 330], [45, с. 187—194], [54, с. 468—469; русский перевод с. 394—395], [91, с. 173—180; русский перевод с. 172—179], [105, с. 236—240].

12.32. РАВЕНСТВО (НЕРАВЕНСТВО) ДВУХ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

Обычная форма записи данных. Заданы две выборки, как в параграфе 12.31. Объем первой выборки равен n_1, второй выборки — n_2. Коэффициент корреляции [8] первой выборки равен r_1, второй выборки – r_2.

Статистическая модель. Выборки производятся из двумерных нормальных совокупностей [9] с коэффициентами корреляции, равными ρ_1 и ρ_2 соответственно. Наблюдения независимы [10].

Гипотезы.

а) *Равенство*	*б) Неравенство*	в) *Неравенство*
$H_0: \rho_1 = \rho_2;$	$H_0: \rho_1 \geqslant \rho_2;$	$H_0: \rho_1 \leqslant \rho_2;$
$H_1: \rho_1 \neq \rho_2.$	$H_1: \rho_1 < \rho_2.$	$H_1: \rho_1 > \rho_2.$

Критическая область.

а) *Равенство.* Выше верхней 2,5 %-ной точки и ниже нижней 2,5 %-ной точки стандартного нормального распределения.

б) *Неравенство.* Ниже нижней 5 %-ной точки стандартного нормального распределения.

в) *Неравенство.* Выше верхней 5 %-ной точки стандартного нормального распределения.

---

[8] Коэффициенты корреляции для выборки и для совокупности (r и ρ соответственно) определены в параграфе 8.8.

[9] Двумерное нормальное распределение рассмотрено в параграфе 9.6.

[10] Каждое наблюдение представляет собой точку с двумя координатами X и Y. Предполагается отсутствие зависимости между парами. Значения координат x и y внутри пары не предполагаются независимыми.

Вычисление значения критериальной статистики. Положим

$$z_1 = \frac{1}{2}\ln\{(1+r_1)/(1-r_1)\},$$

$$z_2 = \frac{1}{2}\ln\{(1+r_2)/(1-r_2)\},$$

$$S^2 = (n_1-3)^{-1} + (n_2-3)^{-1}.$$

Значение критериальной статистики равно:

$$T = (z_1 - z_2)/S. \qquad (12.32.1)$$

Комментарии. 1. Критерий является приближенным. Он основывается на z-преобразовании Фишера (см. табл. 12.31.1).

2. Устойчивость z-преобразования обсуждается в п. 3 комментариев из параграфа 12.31.

Пример 12.32.1. Табл. 12.32.1 содержит данные о весе и росте десяти мужчин и восьми женщин. Требуется проверить гипотезу, состоящую в том, что коэффициент корреляции между ростом и весом для мужчин и женщин совпадает.

Итак, следует проверить:

$$H_0: \rho_м = \rho_ж;$$

$$H_1: \rho_м \neq \rho_ж.$$

Таблица **12.32.1. Вес и рост десяти мужчин и восьми женщин, случайно выбранных из некоторой группы людей**

Мужчины		Женщины	
рост, см	вес, кг	рост, см	вес, кг
166	63	161	60
187	98	158	61
170	82	165	68
171	80	154	52
182	83	165	67
188	85	170	68
175	84	167	69
183	79	177	80
179	75		
178	75		

5 %-ная критическая область состоит из верхней и нижней 2,5 %-ных областей стандартного нормального распределения, что дает $|T| > 1{,}96$. Находим, что

$$n_м = 10, \quad r_м = 0{,}6744, \quad z_м = 0{,}8188;$$

$$n_ж = 8, \quad r_ж = 0{,}9665, \quad z_ж = 2{,}0362.$$

$S^2 = 0{,}342\,86$; значение критериальной статистики равно:

$$T = (0{,}8188 - 2{,}0362)/(0{,}342\,86)^{\frac{1}{2}} = -2{,}08.$$

Эта величина лежит в критической области, поэтому мы отклоняем нулевую гипотезу и заключаем, что $\rho_м \neq \rho_ж$.

Литература: [9, с. 331], [105, с. 241].

12.33. РАВЕНСТВО НЕСКОЛЬКИХ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

Обычная форма записи данных. Заданы k выборок, каждая из которых имеет тот же вид, что и выборки в параграфе 12.31. Объем первой выборки равен n_1, . . ., k-й выборки — n_k. Коэффициент корреляции [11] для первой выборки равен r_1, . . ., для k-й — r_k.

Статистическая модель. Предполагается, что все выборки извлекаются из двумерных нормальных совокупностей [12]; при этом первая выборка — из первой совокупности, вторая — из второй, ..., k-я выборка — из k-й совокупности. Соответствующие коэффициенты корреляции для совокупностей равны ρ_1, ρ_2, . . ., ρ_k. Все наблюдения независимы [13].

Гипотезы.

$$H_0: \rho_1 = p2 = \cdot\ \cdot\ \cdot = рь$$

H_1: нулевая гипотеза неверна.

Критическая область. Верхняя 5 %-ная область распределения χ^2_{k-1}.

Вычисление значения критериальной статистики. Положим

$$z_1 = \frac{1}{2}\ln\{(1+r_1)/(1-r_1)\},$$

.

.

.

$$z_k = \frac{1}{2}\ln\{(1+r_k)/(1-r_k)\},$$

$$N = (n_1 + n_2 + \ldots + n_k) - 3k,$$

$$\mathbf{2} = \{(n_1 - \mathbf{3})\, z_1 + \ldots + (n_k - \mathbf{3})\, z_k\}/N.$$

Значение критериальной статистики равно:

$$\chi^2 = \{(n_1 - 3)\, 2? + \ldots + (n_k - 3)\, z_k^2\} - N\bar{z}^2. \qquad (12.33.1)$$

Комментарии. 1. Критерий является приближенным, при его вычислении используют г-преобразование Фишера.

2. Устойчивость z-преобразования обсуждается в п. 3 комментариев из параграфа 12.31.

3. В случае $k = 2$ критерий эквивалентен критерию а) из параграфа 12.32; величина (12.33.1) есть квадрат величины (12.32.1).

[11] Коэффициенты корреляции для совокупности и выборки определены в параграфе 8.8.

[12] Двумерное нормальное распределение рассмотрено в параграфе 9.6.

[13] См. сноску 10.

Пример 12.33.1. В табл. 12.32.1 приведены данные о росте и весе десяти мужчин и восьми женщин. Применим критерий (12.33.1) для проверки нулевой гипотезы о том, что коэффициенты корреляции для мужчин и женщин одинаковы.

5 %-ная критическая область здесь соответствует верхней 5 %-ной области распределения χ^2_1; она определяется неравенством $\chi^2 > 3{,}84$. Легко найти, что

$$n_м = 10, \quad r_м = 0{,}6744, \quad z_м = 0{,}8188;$$

$$«ж = 8, \quad r_ж = 0{,}9665, \quad z_ж = 2{,}0362.$$

Далее,

$$N = 10 + 8 - 3 \times 2 = 12,$$

$$z = (7 \times 0{,}8188 + 5 \times 2{,}0362)/12 = 1{,}3261.$$

Значение критериальной статистики равно:

$$\chi^2 = 7 \times (0{,}8188)^2 + 5 \times (2{,}0362)^2 - 12 \times (1{,}3261)^2 = 4{,}32.$$

Эта величина лежит в критической области, поэтому мы отклоняем нулевую гипотезу и заключаем, что $\rho_м \neq \rho_ж$.

Заметим, что в этом случае $k = 2$ к значение критерия (4,32) равно квадрату значения критерия из примера 12.32.1 $(-2{,}08)$.

Литература: [9, с. 331], [105, с. 242].

12.34. НЕПАРАМЕТРИЧЕСКИЙ КРИТЕРИЙ НЕКОРРЕЛИРОВАННОСТИ. КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЭНА r_S

Если необходимо исследовать двумерное распределение, заметно отличающееся от нормального, то для получения вывода о коррелированности его компонент нельзя применять критерии проверки гипотез о значении коэффициентов корреляции, рассмотренные в параграфах 12.31, 12.32 и 12.33. Для получения двумерной совокупности, близкой к нормальной, можно прибегнуть к различным преобразованиям данных. В качестве альтернативы может быть предложен способ, основанный на использовании рангов измерений каждой переменной. Этот подход приводит к определению коэффициента ранговой корреляции Спирмэна r_S. Разумеется, этот метод целесообразен также в случае, когда данные с самого начала заданы в виде рангов.

Коэффициент ранговой корреляции Спирмэна r_S для выборки $(x_1, y_1), \ldots, (x_n, y_n)$ объема n определяется как обыкновенный коэффициент корреляции для ранговых переменных. Это означает, что мы ранжируем значения x по возрастанию, приписывая им ранги от 1 до n. Аналогичная операция производится и с y. Затем вычисляем коэффициент корреляции между рангами каждого элемента пары (x_i, y_i) $(i = 1, \ldots, n)$.

Можно воспользоваться вычислительным методом из параграфа 8.8, однако более целесообразно применять формулу (12.34.2), поскольку она более простая. Пусть число элементов

в выборке стремится к бесконечности, тогда в пределе получим генеральный коэффициент ранговой корреляции ρ_S. Выясним, какое соотношение между ρ_S и обычным коэффициентом корреляции ρ двух случайных величин X и Y. Напомним, что функция распределения $F(x)$ случайной переменной задает вероятность того, что $X \leqslant x$ (см. параграф 8.2). Аналогично функция распределения $G(y)$ задает вероятность того, что $Y \leqslant y$. Определим с помощью $F(x)$ и $G(y)$ следующее преобразование переменных: пусть $U = F(X)$, $V = G(Y)$. Ранговый коэффициент корреляции X и У равен обыкновенному коэффициенту корреляции U и V. С помощью этого результата можно показать, что если X и Y имеют двумерное нормальное распределение с коэффициентом корреляции ρ, то

$$\rho_S = \frac{6}{\pi} \arcsin \frac{\rho}{2} = \frac{3}{\pi}\left(\rho + \frac{\rho^3}{24} + \frac{3\rho^5}{640} + \ldots\right). \qquad (12.34.1)$$

В действительности в этом случае ρ_S и ρ приблизительно равны.

Теперь перейдем к рассмотрению критерия некоррелированности.

Обычная форма записи данных. Имеются n точек:

$$(x_1, y_1), (x_2, y_2) \ldots\ldots (x_n, y_n).$$

Статистическая модель. Выборки производятся из двумерной совокупности. Наблюдения независимы.

Гипотезы.

$$H_0: \rho_S = 0;$$

$$H_1: \rho_S \neq 0.$$

Критическая область.

а) *Малые выборки.* Верхняя и нижняя 2,5 %-ные области распределения коэффициента ранговой корреляции Спирмэна (см. табл. 12.34.1).

б) *Большие выборки.* Верхняя и нижняя 2,5 %-ные области t_{n-2}-распределения.

Вычисление значения критериальной статистики.

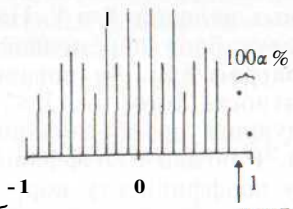
а) *Малые выборки.* Ранжируем значения x в возрастающем порядке от 1 до n. Ранжируем значения y в возрастающем порядке от 1 до n. Вычислим разность между значениями рангов каждого значения x и соответствующего значения y. Обозначим эти разности через $d_1, d_2, \ldots, d_n$. В качестве критерия возьмем коэффициент ранговой корреляции Спирмэна:

$$r_S = 1 - 6(d_1^2 + \ldots + d_n^2)/(n^3 - n). \qquad (12.34.2)$$

б) *Большие выборки.* Вычислим r_S, как в п. а). Затем используем в качестве статистического критерия

$$t = (n-2)^{\frac{1}{2}} r_S / (1 - r_S^2)^{\frac{1}{2}}. \qquad (12.34.3)$$

Таблица 12.34.1, **Верхние * 100α %-ные точки распределения коэффициента ранговой корреляции Спирмэна (критическая область включает табличное значение критерия)**



Табличное значение критерия

Объем выборки n -	a = 0,05	0,025	0,01
5	0,900	1,000	1,000
6	0,828	0,885	0,942
7	0,714	0,785	0,892
8	0,642	0,738	0,833
9	0,600	0,683	0,783
10	0,563	0,648	0,745
11	0,527	0,609	0,700
12	0,497	0,580	0,671
13	0,478	0,555	0,643
14	0,459	0,534	0,622
15	0,443	0,518	0,600
16	0,427	0,500	0,582
17	0,412	0,485	0,564
18	0,399	0,472	0,548
19	0,390	0,458	0,533
20	0,379	0,445	0,520
21	0,369	0,435	0,508
22	0,360	0,424	0,496
23	0,352	0,415	0,485
24	0,344	0,406	0,475
25	0,336	0,398	0,465
26	0,330	0,389	0,456
27	0,324	0,382	0,448
28	0,318	0,375	0,440
29	0,311	0,369	0,432
30	0,306	0,362	0,425

* Нижняя 100α %-ная точка получается при изменении знака верхней 100α %-ной точки на противоположный.

Источники. Элементы таблицы для га^П перепечатаны из [34] с разрешения издателя. Для $n<11$ они вычислены с помощью таблицы точного распределения $\sum d_i^2$ из [74].

Комментарий. Если два или более значения совпадают, то им приписываем ранг, равный среднему рангов, приписываемых при отсутствии совпадения.

Пример 12.34.1. В табл. 12.31.2 приведены оценки пятнадцати студентов за ответы на два экзаменационных вопроса. Проверим пулевую гипотезу, состоящую в том, что оценки за ответы на вопрос 1 и на вопрос 2 некоррелированы.

Согласно табл. 12.34.1 5 %-ная критическая область есть $|r_S| > 0{,}518$. Для вычисления статистики построим таблицу (см. табл. 12.34.2), в которой показаны ранги. Коэффициент ранговой корреляции равен:

$$r_S = 1 - 6\{0^2 + (10{,}5)^2 + \ldots + 0^2\}/\{(15)^3 - 15\} = 0{,}704.$$

Эта величина значима на 5 %-ном уровне; мы отклоняем нулевую гипотезу и заключаем, что ps $\neq 0$.

Таблица 12.34.2. **Ранжирование оценок студентов за ответы на два экзаменационных вопроса**

Студент	Вопрос 1, ранг	Вопрос 2, ранг	Разность
1	15	15	0
2	12,5	2	10,5
3	8	10	2
4	4	7,5	3,5
5	1	1	0
6	12,5	10	2,5
7	2	5,5	3,5
8	8	7,5	0,5
9	8	10	2
10	5	4	1
11	6	5,5	0,5
12	10,5	14	3,5
13	14	13	1
14	10,5	12	1,5
15	3	3	0

К этим данным был применен второй критерий. Необходимо обратить внимание на предупреждение, сделанное в параграфе 12.2.

Литература: [52, с. 128—129; русский перевод с. 142—143], [81, с. 145—150], [91, с. 190—199; русский перевод с. 178—189], [105, с. 243].

12.35. ПРОВЕРКА НОРМАЛЬНОСТИ

Предположим, что имеется выборка из совокупности, распределение которой неизвестно. Все наблюдения независимы. Необходимо проверить, является ли это распределение нормальным. Для этой цели можно применить критерий согласия χ^2. Мы вычисляем выборочное среднее x, выборочную дисперсию s^2 и берем нормальную кривую с параметрами x и s^2 (таким образом,

$\mu = x$, $\sigma^2 = s^2$). Затем вычисляем ожидаемые числа наблюдений, попадающих в последовательные интервалы в соответствии с выбранной нормальной кривой, и производим обычную проверку подбора распределения по критерию согласия χ^2. Критерием Колмогорова—Смирнова следует пользоваться с осторожностью (см. параграф 12.12).

В книге Снедекора [91, с. 201—203; русский перевод с. 195—197] описаны процедуры проверки гипотез о нормальности с помощью вычисления показателей асимметрии и эксцесса. Следует предпочесть им проверки с помощью критериев согласия, о которых говорилось выше.

Пример 12.35.1. Из совокупности с неизвестным распределением произведена выборка объема 20. Значения наблюдений следующие:

8,2	3,9	2,9	0,0	1,2	2,4	**0,1**
4,2	7,7	9,8	3,5	8,5	**1,3**	7,7
8,6	3,8	3,9	0,7	4,1	6,0	

Требуется проверить нулевую гипотезу, состоящую в том, что распределение совокупности нормально.

Для начала следует вычислить значения выборочного среднего $x = 4{,}425$ и выборочного стандартного отклонения s = 3,093. Согласно подобранному нормальному распределению со средним $\mu = 4{,}425$ и стандартным отклонением $\sigma = 3{,}093$ ожидаемое число исходов с величиной, лежащей в интервале от $\mu + 0{,}5\sigma$ до $\mu + 1{,}5\sigma$, равно:

$$20\,(\Phi\,(1{,}5) - \Phi\,(0{,}5)) = 4{,}8346.$$

Аналогичным образом вычисляются ожидаемые числа исходов для других интервалов в табл. 12.35.1 (заметим, что интервалы в табл. 12.35.1 выбраны из соображений удобства, можно было бы работать с другими интервалами). Значение χ^2 оказывается равным 8,44. Оно лежит в критической области, которая совпадает с верхней 5 %-ной областью распределения χ^2 при трех степенях

Таблица 12.35.1. **Метод χ^2 при проверке нормальности распределения**

Интервал	Число исходов		Вклад в χ^2
	наблюдаемое	ожидаемое	
От $-\infty$ до $\mu - 1{,}5\sigma$	0	1,3362	1,34
От $\mu - 1{,}5\sigma$ до $\mu - 0{,}5\sigma$	6	4,8345	0,28
От $\mu - 0{,}5\sigma$ до μ	7	3,8292	2,63
От μ до $\mu + 0{,}5\sigma$	0	3,8292	3,83
От $\mu + 0{,}5\sigma$ до $\mu + 1{,}5\sigma$	6	4,8345	0,28
От $\mu + 1{,}5\sigma$ до ∞	1	1,3362	0,08
Всего	20	20,0000	8,44

свободы. Поэтому следует отклонить нулевую гипотезу и заключить, что распределение совокупности не нормально.

Литература: [3, с. 225], [9, с. 394], [64], [86, с. 218—224, 233], [91, с. 201—202; русский перевод с. 195—197], [105, с. 79—85].

12.36. ПРОВЕРКА ГИПОТЕЗЫ О ТОМ, ЧТО СОВОКУПНОСТЬ РАСПРЕДЕЛЕНА ПО ЗАКОНУ ПУАССОНА

Пусть имеется выборка независимых наблюдений $x_1, \ldots, x_n$ из неизвестного распределения, причем Xi — целые неотрицательные числа. Пусть также требуется проверить, является ли пуассоновской совокупность, из которой извлечена выборка. Для этой цели можно применить критерий согласия χ^2 из параграфа 12.11. Следует вычислить выборочное среднее x и рассмотреть распределение Пуассона со средним (являющимся единственным его параметром) $\lambda = x$. Далее нужно вычислить число ожидаемых исходов в последовательных интервалах, основываясь на подобранном распределении. Затем необходимо провести обычную процедуру проверки подбора распределения по критерию согласия χ^2. Критерий Колмогорова—Смирнова должен применяться с осторожностью (см. параграф 12.12).

Заметим, что среднее и дисперсия распределения Пуассона совпадают. Если выборочное среднее и выборочная дисперсия s^2 заметно различаются, то можно сомневаться в применимости модели Пуассона. Прежде чем выполнить всю процедуру выбора пуассоновского распределения и проверки правильности выбора, имеет смысл применить следующий предварительный критерий проверки равенства среднего и дисперсии.

Обычная форма записи данных. $x\backslash, \ldots, x_n$ — выборка объема n. Выборочные среднее и дисперсия равны x *я* s^2 соответственно. Числа $\{x_i\}$ — целые неотрицательные.

Статистическая модель. Числа $\{x_i\}$ — независимые наблюдения, выбранные из унимодального распределения на множестве неотрицательных целых чисел. Распределение имеет среднее μ и дисперсию σ^2, оно близко к нормальному.

Гипотезы.

$$H_0 : \sigma^2 = \mu;$$
$$H_1 : \sigma^2 \neq \mu.$$

Критическая область. Верхняя 2,5 %-ная и нижняя 2,5 %-ная области распределения χ^2 с $(n - 2)$ степенями свободы.

Вычисление значения критериальной статистики. Значение критериальной статистики равно:

$$\chi^2 = (n - 1)\, s^2/\bar{x}. \qquad (12.36.1)$$

Комментарии. 1. Этот критерий является приближенным. Его не следует применять, если nx меньше 10.

2. Критерий основывается на формуле (12.28.1). Неизвестная выборочная дисперсия σ^2 заменяется выборочным средним x, которое дает оценку дисперсии распределения Пуассона.

3. Критерий предполагает, что распределение унимодальное и близко к нормальному; он проверяет величину дисперсии, а не форму распределения. Нулевая гипотеза о распределении по закону Пуассона должна быть отклонена, если величина критерия значима. В противном случае должна быть проведена полная проверка подбора распределения по критерию согласия. Нулевая гипотеза о пуассоновском распределении отклоняется при значимой величине критерия.

Литература: [13], [64], [86, с. 246—248], [105, с. 302—305].

12.37. ПРОВЕРКА ДРУГИХ ТИПОВ РАСПРЕДЕЛЕНИИ

В параграфах 12.35 и 12.36 мы показали, каким образом критерий согласия χ^2 может применяться для проверки гипотез о нормальном и пуассоновском распределениях. Та же процедура может использоваться для распределений другой формы. Мы оцениваем параметры распределения с помощью методов из главы 13, а затем вычисляем ожидаемые числа наблюдений в последовательных интервалах. Затем следует обычным способом применить критерий согласия χ^2. При применении метода Колмогорова—Смирнова необходимо соблюдать осторожность (см. параграф 12.12).

12.38. КРИТЕРИИ, ИСПОЛЬЗУЮЩИЕ ПУАССОНОВСКИЕ ПЕРЕМЕННЫЕ

Единственным параметром распределения Пуассона является его среднее значение Я. Поэтому гипотезы о распределениях Пуассона сводятся к гипотезам об их средних значениях. Для их проверки могут применяться критерии, основанные на нормальном распределении, при условии, что выборки не слишком малы, а средние значения не меньше девяти или десяти. Некоторые из таких критериев предполагают равенство дисперсий (например, дисперсионный анализ по одному признаку из параграфа 12.18). Для выполнения этого требования может применяться описанное в параграфе 14.6 преобразование квадратного корня. Это преобразование дает также распределение, более близкое к нормальному, чем первоначальное.

12.39. УПРАЖНЕНИЯ

1. Предприятие решает приобрести один из двух станков. Более дорогой станок A по утверждению продавца дает три бракованные детали из сотни. В процессе испытания на станке A произведено 15 бракованных деталей из общего числа, равного 120, в то время как на станке B — 14 бракованных деталей из 130. Противоречат ли эти результаты утверждению о качестве работы станков?

2. Существует ли зависимость в следующей таблице сопряженности признаков?

5	2	7
3	8	11
8	10	18

3. Проверьте подбор распределения по логарифмическому ряду в примере из параграфа 10.12.2.

4. Проверьте нулевую гипотезу о том, что среднее совокупности из примера 12.12.1 равно нулю, предполагая, что эта совокупность нормальна.

5. Проверьте нулевую гипотезу о том, что дисперсия совокупности из примера 12.12.1 равна 1, предполагая, что эта совокупность нормальна.

6. Следующие две выборки извлечены из совокупностей, имеющих пуассоновское распределение:

14, 7, 13, 8, 5, 10, 5, 4, 7, 16, 11, 11, 6, 12, 10;
12, 7, 7, 7, 10, 8, 13, 7, 11, 14, 9, 7, 8, 6.

Верно ли, что совокупности имеют одинаковые средние?

7. Примените процедуру дисперсионного анализа по одному признаку к задаче о содержании холестерола у черепах из примера 12.15.2. Проверьте, что полученное значение F равно квадрату значения t из примера 12.15.2.

8. Допустим, что верхнее левое значение в табл. 12.25.3 пропущено. Проверьте нулевую гипотезу, состоящую в том, что в содержании фосфора у животных трех пород различия нет.

9. Проверьте нулевую гипотезу, состоящую в том, что отметки в таблице 12.31.2 независимы.

13. ТОЧЕЧНОЕ И ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

В первом параграфе этой главы изучаются общие принципы точечного оценивания, метод максимального правдоподобия, построение доверительных интервалов. Эти методы далее применяются для оценивания параметров в дискретном распределении общего вида (параграфы 13.5—13.11), оценивания среднего и линейной комбинации средних нормальной генеральной совокупности (параграфы 13.12—13.18), оценивания дисперсий и отношения дисперсий нормальных совокупностей (параграфы 13.19—13.24), оценивания параметров логарифмически-нормального распределения (параграфы 13.25) и оценивания коэффициента корреляции в двумерном нормальном распределении (параграф 13.26). В параграфе 13.27 описывается метод построения доверительных границ функции распределения.

13.1. ТОЧЕЧНОЕ ОЦЕНИВАНИЕ

Рассмотрим выборку, состоящую из n наблюдений: $x_1, \ldots, x_n$. Форма распределения генеральной совокупности известна (например, нормальная), однако один или несколько параметров неизвестны (например, μ, σ^2 либо одновременно μ и σ^2). Задача заключается в том, чтобы оценить неизвестные параметры.

Функция наблюдений, которую мы используем для измерения параметров, называется *оценкой* (estimator), а ее численное значение, полученное на основе имеющегося множества данных, назовем *значением оценки* (estimate).

В качестве оценки часто можно предложить несколько функции (например, выборочное среднее и выборочная медиана могут одновременно служить оценкой среднего нормальной совокупно-

сти), и мы должны решить, какую из них выбрать. Статистики-исследователи при этом руководствуются следующим:

1. Оценка должна быть *несмещенной,* т. е. ее математическое ожидание равно истинному значению параметра. Таким образом, в среднем значения оценки будут равны оцениваемому параметру. Оценка не должна приводить к значениям, которые в среднем либо слишком велики, либо слишком малы.

2. Оценка должна быть *состоятельной.* Под этим понимается следующее: для любой достаточно малой величины ε вероятность того, что абсолютная величина отклонения оценки от истинного значения меньше ε, стремится к 1 при неограниченном увеличении п. Таким образом, для оценки, построенной на основе большого числа наблюдений, вероятность того, что значение оценки будет отличаться от истинного значения параметра, весьма мала.

3. Оценка должна быть *эффективной.* Это означает, что дисперсия такой оценки минимальна.

Литература: [2, с. 21—26; русский перевод с. 35—39], [7, с. 88—91; русский перевод с. 78—80], [26, с. 214—224], [45, с. 137—138], [46, с. 190—196], [81, с. 151—156], [86, с. 148—151], [93, с. 156—166].

13.2. МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Метод максимального правдоподобия приводит к оценкам, которые обычно весьма удовлетворительны. Эти оценки имеют такие желательные свойства, как состоятельность, асимптотическая нормальность и эффективность для больших выборок при самых общих условиях. Они, как правило, смещены, однако смещение часто можно устранить простым способом (см. пример 13.2.2). Существуют и другие методы получения оценок, однако метод максимального правдоподобия применяется наиболее часто.

Оценки по методу максимального правдоподобия имеют другое, весьма полезное свойство — *инвариантность.* Обозначим оценку максимального правдоподобия параметра б через $\widehat{\theta}$[1]. Тогда, если $f(\theta)$ — взаимно-однозначная функция 0, то оценкой максимального правдоподобия $f(\theta)$ будет $f(\widehat{\theta})$. Например, $\widehat{\sigma} = (\widehat{\sigma^2})^{1/2}$.

По *методу максимального правдоподобия* мы должны выбрать оценку, которая максимизирует правдоподобие наблюдаемого события[2]. Следующие примеры иллюстрируют описанный метод.

Пример 13.2.1. Найдем оценку максимального правдоподобия $\widehat{p}$ вероятности p в эксперименте с подбрасыванием монеты, если известно, что в n экспериментах r раз выпал орел.

---

[1] Такое обозначение общепринято.

[2] Различие между правдоподобием и вероятностью объясняется в [54, сноска на с. 8; русский перевод с.23].

Из (10.1.1) следует, что вероятность наблюдаемого события равна:

$$L = \binom{n}{r} p^r (1-p)^{n-r}.$$

В соответствии с методом максимального правдоподобия мы выбираем такое значение p, которое максимизирует L. Это значение максимизирует также

$$\ln L = \ln \binom{n}{r} + r \ln p + (n-r) \ln (1-p).$$

В точке максимума производная по p должна быть равна 0. Поэтому

$$\frac{\partial}{\partial p} \ln L = r/p - (n-r)/(1-p) = 0$$

и

$$\hat{p} = r/n. \qquad \textbf{(13.2.1)}$$

Пример 13.2.2. Выборка из n наблюдений $x_1, \ldots, x_n$ извлечена из нормальной совокупности. Найдем оценку максимального правдоподобия математического ожидания $\widehat{\mu}$ и дисперсии совокупности $\widehat{\sigma^2}$.

Функция правдоподобия в данном случае равна:

$$L = \left[\frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2\right\}\right] \cdots \left[\frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2\right\}\right].$$

В соответствии с методом максимального правдоподобия мы выбираем такие значения μ и σ^2, которые максимизируют L. Эти же значения максимизируют $\ln L$. Поэтому приравняем частные производные $\widehat{\mu}$ по $\widehat{\sigma^2}$ нулю (см. параграф 1.6). Находим

$$\widehat{\mu} = x \qquad \textbf{(13.2.2)}$$

и

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Последняя оценка имеет небольшое смещение, однако его можно избежать умножением ее на множитель $n/(n-1)$, что приводит к оценке

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad (13.2.3)$$

которая совпадает с выборочной дисперсией (8.4.3).

Литература: [2, с. 21—26; русский перевод с. 35—39], [3, с. 104—106], [7, с. 91—95; русский перевод с. 80—84], [26, с. 224—228], [45, с. 137—155], [65, с. 203], [66, с. 158—159].

13.3. ДРУГИЕ МЕТОДЫ

Метод максимального правдоподобия в некоторых случаях может приводить к большому количеству вычислений, которые часто не оправданы или приводят к затруднениям. Существуют и другие методы, наиболее известный из которых — метод моментов. Его суть заключается в приравнивании нескольких первых выборочных моментов к моментам генеральной совокупности. Этот метод рассмотрен в гл. 11, примеры приведены в параграфах 10.11 и 10.12.

Метод наименьших квадратов разбирается в III части книги.

Литература: [26, с. 242—243], [70, с. 186—187], [92, с. 544].

13.4. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

Методы, описанные в параграфах 13.2 и 13.3, применимы при так называемом *точечном оценивании параметров,* однако следует помнить, что оценка является случайной величиной, распределенной некоторым образом около истинного значения параметров. Истинный параметр может быть меньше или больше нашей оценки. Часто полезно построить интервал, внутри которого истинное значение будет лежать с достаточной вероятностью, границы этого интервала будем называть *доверительными границами.*

Следующая процедура позволяет определить 95 %-ные нижние и верхние доверительные границы. Когда мы говорим, что данный интервал накрывает истинное значение, это означает, что наше утверждение имеет место в 95 случаях из 100 и только в 5 случаях оно неверно.

1. Выбрать некоторую критериальную статистику, зависящую от данного неизвестного параметра, но не зависящую от других неизвестных параметров.
2. Подставить в формулу статистики значения, соответствующие данной выборке.
3. Приравнять критериальную статистику к верхней 2,5 %-ной точке соответствующего распределения.
4. В качестве первой доверительной границы взять корень предыдущего уравнения.
5. Повторить п. 3 для нижней 2,5 %-ной точки и найти вторую доверительную границу.

Предупреждение. Обратите внимание на замечания в параграфе 12.5 относительно пригодности выбранной статистической модели.

Можно строить 95 %-ные интервалы с помощью неравных хвостов распределения (например, выбрать 2 %-ную и 3 %-ную точки). Доверительные интервалы должны быть как можно более

короткими; это условие достигается при использовании симметричных распределений типа нормального или t-распределения с равными хвостами. Та же процедура весьма незначительно минимизирует доверительный интервал при других, несимметричных распределениях (например, распределение χ^2), в этих случаях можно избежать довольно утомительных вычислений.

Если соответствующая статистика содержит квадрат неизвестного параметра, то *обе* доверительные границы получаются приравниванием этой статистики к верхней 5 %-ной точке соответствующего распределения. (Примеры приведены в параграфах 13.7, 13.15, 13.18.) Применение двух хвостов в этом случае приводит к паре непересекающихся интервалов.

Если имеется два или большее число параметров, то возможно построение области, для которой с определенной уверенностью мы можем утверждать, что истинный параметр лежит внутри нее. Такие области называются *доверительными областями.* Так, в примере с триномиальным экспериментом из параграфа 13.7 нами получена область, внутри которой с вероятностью 95 % *одновременно* находятся параметры p_1, p_2 и p_3. Однако интервал для p_1 не будет 95 %-ным. Неверно также и то, что 85,7375 %-ную доверительную область для p_1, p_2 и p_3 можно получить пересечением трех различных 95 %-ных доверительных интервалов, поскольку статистики, используемые при построении индивидуальных доверительных интервалов, не являются независимыми. Это обычная проблема многомерного распределения с большим числом параметров, с ней сталкиваются и в случае нормального распределения, поскольку статистика (12.13.1), применяемая при построении доверительного интервала для среднего, и статистика (12.28.1), необходимая для доверительного интервала дисперсии, не являются независимыми. Указанные проблемы возникают только при большом числе параметров.

Пример 13.4.1. Из нормального распределения с неизвестными средним и дисперсией извлечена выборка, содержащая 9 единиц. Выборочное среднее равно 4,2, выборочная дисперсия равна 1,69. Найдем 95 %-ный интервал для среднего μ.

Соответствующая статистика для этой задачи определяется с помощью выражения (12.13.1). Доверительные границы находят из решения уравнений

$$\text{У9}\ (4{,}2 - \mu)/(1{,}69)^{\frac{1}{2}} = 2{,}306,$$

$$\sqrt{9}\ (4{,}2 - \mu)/(1{,}69)^{\frac{1}{2}} = -2{,}306.$$

% %-ный доверительный интервал заключен между 3,2 и 5,2.

Литература: [2, с. 21—26], [7, с. 121—127; русский перевод С. 114—120], [26, с. 276—289], [46, с. 200—202], [70, с. 248—270], [81, с. 151—156], [86, с. 152—187], [93, с. 156—166], [102, с. 195—215; русский перевод с. 374—391].

13.5. БИНОМИАЛЬНЫЙ ПАРАМЕТР p

Обозначение и статистическая модель. См. параграф 12.6. В выборке объема n r обозначает количество «успехов».

Точечная оценка.

$$p = r/n. \tag{13.5.1}$$

95 %-ные доверительные границы. Приравняем (13.5.2) к верхней и нижней 2,5 %-ным точкам F-распределения с $2(r+1)$ и $2(n-r)$ степенями свободы:

$$p = \left(\frac{n-r}{r+1}\right)\left(\frac{p}{1-p}\right). \tag{13.5.2}$$

Комментарии. 1. Связь биномиального и F-распределения была объяснена в параграфе 10.3. Метод нахождения нижних процентных точек F-распределения был описан в примере 9.4.3.

2. При большом n (например, при $n > 20$) биномиальное распределение можно аппроксимировать нормальным. В этом случае (12.6.1) необходимо приравнять к верхней и нижней 2,5%-ным точкам нормального распределения.

3. Можно также воспользоваться биномиальными таблицами (см. параграф 10.1). Рассмотрим таблицу, соответствующую данному n при различных значениях p. Сначала найдем такое значение p, для которого верхняя 2,5 %-ная точка равна r. Таким образом получим одну доверительную границу. Затем выберем другое значение p, для которого нижняя 2,5 %-ная точка равна r. Последнее значение примем в качестве второй доверительной границы.

4. Таблицы, в которых даны доверительные границы для p, приведены в [76, с. 204] *.

5. Задача оценивания параметра усеченного биномиального распределения рассматривается в [47, с. 73—76].

Пример 13.5.1. В биномиальном эксперименте было произведено 18 испытаний, 14 из которых оказались успешными. Оценим p (вероятность успеха) и найдем 95 %-ный доверительный интервал для p.

Как следует из (13.5.1), точечной оценкой p является $\hat{p} = {}^{14}/_{18} = 0{,}78$. Для нахождения 95 %-ных доверительных границ приравняем $\frac{4}{15}\frac{\hat{p}}{p}/(1-p)$ к верхней и нижней 2,5 %-ным точкам $F_{30,8}$-распределения:

$$\frac{4}{15}p/(1-p) = 3{,}89,$$

$$\frac{4}{15}p/(1-p) = 1/2{,}65.$$

Доверительными границами будут 0,586 и 0,936.

* См. также: Янко Я. Математико-статистические таблицы. М., Госстатиздат, 1961.— *Примеч. пер.*

Для построения доверительного интервала с помощью аппроксимации нормальным распределением необходимо решить следующие уравнения:

$$(14 - 18p)/\{18p(1-p)\}^{\frac{1}{2}} = \pm 1{,}96.$$

Возведем обе части уравнений в квадрат и домножим их на $18p(1-p)$:

$$(14 - 18p)^2 = (1{,}96)^2\{18p(1-p)\}.$$

После преобразования придем к квадратному уравнению:

$$393{,}1488p^2 - 573{,}1488p + 196 = 0.$$

Корнями этого уравнения будут 0,548 и 0,910. Эти значения и образуют приближенные 95 %-ные доверительные границы.

Литература: [2, с. 21—26; русский перевод с. 35—39], [7, с. 91—92, 129—130; русский перевод с. 99—101, 137—140], [46, с. 196—200], [47, с. 58-59], [70, с. 260-264], [93, с. 156-166], [102, с. 195—200; русский перевод с. 377—378].

13.6. РАЗНОСТЬ ДВУХ БИНОМИАЛЬНЫХ ВЕРОЯТНОСТЕЙ

Обозначение и статистическая модель. Имеются две выборки. Первая выборка объема n_1, r_1 обозначает количество «успехов», вторая выборка объема n_2, r_2 — количество «успехов». Параметрами для первого эксперимента являются n_1 и p_1, для второго — n_2 и p_2. Величины p_1 и p_2 неизвестны. Обозначим $q_1 = 1 - p_1$, $q_2 = 1 - p_2$ и $\delta = p_1 - p_2$.

Точечная оценка.

$$\hat{\delta} = r_1/n_1 - r_2/n_2. \tag{13.6.1}$$

95 %-ные доверительные границы. Вычислим

$$\hat{p}_1 = r_1/n_1,\ \hat{q}_1 = 1 - \hat{p}_1,\ \hat{p}_2 = r_2/n_2,\ \hat{q}_2 = 1 - \hat{p}_2.$$

Приравняем

$$T = (\hat{p}_1 - \hat{p}_2 - \delta)/(\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2)^{\frac{1}{2}} \tag{13.6.2}$$

к верхней и нижней 2,5 %-ным точкам стандартного нормального распределения.

Комментарий. Данный метод построения доверительного интервала основывается на аппроксимации биномиального закона распределения нормальным. Поэтому объемы выборок n_1 и n_2 должны быть больше 20 (см. параграф 10.2).

Литература: [105, с. 296—297].

13.7. БИНОМИАЛЬНЫЕ ПАРАМЕТРЫ $\{p_i\}$

Обозначение и статистическая модель. См. параграфы 10.4 и 12.7. В выборке объема n n_i исходов имеют вид i $(i = 1, \ldots, k)$.

Точечные оценки.

$$fii=tii/n, \ . \ . \ ., \ p_k = n_k/n. \qquad (13.7.1)$$

95 %-ные доверительные границы. При построении доверительных интервалов для отдельного параметра p могут быть применены методы из параграфа 13.5. Для нескольких параметров необходимо определить доверительную область. Эта область получается приравниванием (12.7.1) к верхней 5 %-ной точке распределения χ^2 с k- 1 степенями свободы при ограничении (10.4.2).

Комментарии. 1. Применение метода построения доверительных интервалов, рассмотренного в параграфе 13.5, к каждому отдельному параметру p (см. параграф 13.4) некорректно. Процедуры построения одновременных доверительных интервалов для *Pi,* p_2 и *рз* описаны в [47, с. 289].

2. Необходимо иметь в виду п. 1 комментариев из параграфа 12.7.

Пример 13.7.1. В триномиальном эксперименте из 50 испытаний 10 результатов имеют вид 1, 24 результата — вид 2 и 16 — вид 3. Найдем оценки максимального правдоподобия параметров *pi, p2, рз* и определим 95 %-ные доверительные границы для параметра p_1, а также 95 %-ную доверительную область для p_1, p_2, *рз.*

Оценками максимального правдоподобия для p_1, p_2 и p_3 являются $^{10}/_{50} = 0{,}2$; $^{24}/_{50} = 0{,}48$ и $^{16}/_{50} = 0{,}32$ соответственно. Для получения 95 %-ных доверительных границ для p_1 решим следующие уравнения:

$$\frac{40}{-} p_1/(1 - p_1) = 1{,}87, \qquad \frac{40}{11} p_1/(1 - p_1) = 1/2{,}12.$$

Нижней и верхней доверительными границами будут 0,115 и 0,340.

Границы 95 %-ной доверительной области для p_1, p_2, *рз* задают уравнения

$$(10 - 50p_1)^2/50p_1 + (24 - 50p_2)^2/50p_2 + (16 - 50p_3)^2/50p_3 = 5{,}99,$$

$$p_1 + P2 + p_3 = 1.$$

Переменную *рз* можно исключить из первого уравнения, выразив ее как $1 - p_1 - p_2$. Полученная таким образом доверительная область для p_1 и p_2 не будет иметь форму эллипса. Однако можно воспользоваться приближением, заменив математические ожидания знаменателей наблюдаемыми значениями (оценками), после чего придем к эллипсу:

$$(10 - 50p_1)^2/10 + (24 - 50p_2)^2/24 + (50p_1 + 50p_2 - 34)^2/16 = 5{,}99.$$

Эта приближенная доверительная область показана на рис. 13.7.1.

Литература: [47, с. 288—290].

Рис. 13.7.1. Приближенная 95 %-ная доверительная область для *р* и *P2*

13.8. ПУАССОНОВСКИЙ ПАРАМЕТР λ

Обозначение и статистическая модель. См. параграф 10.6. Выборка объема п, средняя равна $\bar{x}$.

Точечная оценка,

$$\mathbf{A} = X. \qquad (13.8.1)$$

95 %-ные доверительные границы. а) *пх велико* (например, больше 10). Приравняем

$$T = n^{\frac{1}{2}} (\bar{x} - \lambda)/\lambda^{\frac{1}{2}} \qquad (13.8.2)$$

к верхней и нижней 2,5 %-ным точкам стандартного нормального распределения.

б) *пх мало.* Приравняем $2n\lambda$ к верхней и нижней 2,5 %-ным точкам распределения χ^2 с $2(n\bar{x}+1)$ степенями свободы.

Комментарии. 1. Связь между пуассоновским и распределением χ^2 объяснена в параграфе 10.9.

2. Формула (13.8.2) дает возможность прибегнуть к аппроксимации распределения Пуассона нормальным (см. параграф 10.8).

3. Альтернативный подход: воспользоваться таблицами пуассоновского распределения (см. параграф 10.6). Рассмотрим таблицы, соответствующие различным значениям параметра пуассоновского распределения. Найдем табличное значение, для которого верхняя 2,5 %-ная точка совпадает с *пх.* Разделим найденное значение на *п.* Это даст нам первую доверительную границу. Вторая доверительная граница получается выбором табличного значения, для которого нижняя 2,5 %-ная точка равна nx, после чего последняя делится на *п.*

4. В [76, с. 203] приведена таблица, по которой доверительные границы находятся непосредственно *.

* Доверительные границы для $n\lambda$ приводятся в работе: Хан Г., Шапиро С. Статистические модели в инженерных задачах. М., Мир, 1969, с. 188—189.— *Примеч. ред.*

5. Задача оценивания параметра усеченного пуассоновского распределения обсуждается в [47, с. 104—109]. Приближенный метод приводится в параграфе 14.12.

Пример 13.8.1. Южная часть Лондона была разделена на 576 небольших участков одинаковой площади. Кларк [12] утверждает, что во время второй мировой войны на 229 участков при бомбежке города не попало ни одной бомбы, на 211 участков попало по одной бомбе, на 93 участка — по две бомбы, на 35 — по три бомбы, на 7 — по четыре бомбы и на один участок — пять и более бомб. Найдем оценку максимального правдоподобия параметра пуассоновского распределения K (среднее число попаданий на один участок). Вычислим также 95 %-ный доверительный интервал для λ.

Выборочное среднее равно 0,929, которое и будет оценкой максимального правдоподобия параметра K. 95 %-ные доверительные границы получаются из решения уравнений

$$\sqrt{576}\,(0{,}929-\lambda)/\lambda^{\frac{1}{2}} = \pm 1{,}96.$$

Для решения уравнений возведем их в квадрат и домножим на А. Получим квадратное уравнение

$$576\lambda^2 - 1074{,}0496\lambda + 497{,}1116 = 0$$

с корнями 0,854 и 1,011. Найденные значения составляют 95 %-ные доверительные границы параметра K.

Литература: [2, с. 21—26; русский перевод с. 35—39], [47, с. 94—98, 104—109], [66, с. 160—161].

13.9. ПАРАМЕТР p ГЕОМЕТРИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

Обозначение и статистическая модель. См. параграф 10.10. Объем выборки равен n, выборочное среднее $\bar{x}$.

Точечная оценка.

$$\hat{p} = 1/(1+\bar{x}). \qquad (13.9.1)$$

95 %-ные доверительные границы. а) *Большие выборки* (например, $n > 10$). Приравняем

$$T = (np\bar{x} - nq)/(nq)^{\frac{1}{2}} \qquad (13.9.2)$$

к верхней и нижней 2,5 %-ным точкам стандартного нормального распределения.

б) *Малые выборки.* Воспользуемся таблицами отрицательного биномиального распределения (см. параграф 10.11). Рассмотрим различные таблицы, в которых параметр k равен n. Выберем табличное значение, для которого верхняя 2,5 %-ная точка равна $n\bar{x}$; параметр p этого значения задаст одну доверительную границу. Вторая доверительная граница получается выбором табличного значения, для которого нижняя 2,5 %-ная точка равна $n\bar{x}$.

Комментарии. 1. Оценка максимального правдоподобия (13.9.1) совпадает с оценкой по методу моментов при приравнивании (10.10.2) к величине x.

2. При описании отрицательного биномиального распределения не все авторы пользуются этими же параметрами, поэтому читатель должен быть осторожен.

Пример 13.9.1. Из генеральной совокупности с геометрическим распределением произведена выборка объема 15. Выборочное среднее равно 4,3. Найдем оценку максимального правдоподобия параметра p и его 95 %-ный доверительный интервал.

Оценка максимального правдоподобия равна: $\hat{p} = 1/5{,}3 = 0{,}1875$. Для построения 95 %-ного доверительного интервала параметра p необходимо решить уравнения

$$(65p - 15q)/\sqrt{15q} = \pm 1{,}96.$$

Возведем их в квадрат, домножим на $15q$ и выразим q через $1-p$. После приведения подобных членов получим следующее квадратное уравнение:

$$6400p^2 - 2342{,}376p + 167{,}376 = 0,$$

которое имеет корни 0,097 и 0,269. Эти значения и составляют 95 %-ные доверительные границы параметра p.

13.10. ПАРАМЕТРЫ ОТРИЦАТЕЛЬНОГО БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Обозначение и статистическая модель. См. параграф 10.11. Объем выборки равен n, среднее равно x, выборочная дисперсия — s^2.

Точечные оценки.

$$\hat{p} = \bar{x}/s^2, \quad \hat{k} = \bar{x}\hat{p}/(1-\hat{p}). \qquad (13.10.1)$$

Комментарии. 1. Для рассматриваемого случая простого метода построения доверительных интервалов не существует.

2. Оценки (13.10.1) основаны на методе моментов. Возможно также применение метода максимального правдоподобия, однако этот метод приводит к громоздким вычислениям (см., например, [47, с. 131—135]).

3. Проблема оценивания параметров усеченного отрицательного распределения обсуждается в [47, с. 136—137]. Пример приведен в параграфе 10.11.

Литература: [47, с. 131—135].

13.11. ПАРАМЕТР a РАСПРЕДЕЛЕНИЯ ПО ЛОГАРИФМИЧЕСКОМУ РЯДУ

Обозначение и статистическая модель. См. параграф 10.12. Объем выборки равен n, среднее равно x.

Точечная оценка.

$$\{\hat{a}/(1-\hat{a})\}/\{-\ln(1-\hat{a})\} = \bar{x}. \qquad (13.11.1)$$

Комментарии. 1. Простых методов построения доверительного интервала не существует.

2. Для нахождения точечной оценки параметра a необходимо решить уравнение (13.11.1) либо итеративно, либо методом проб и ошибок.

3. Уравнение (13.11.1) приводит к оценке по методу максимального правдоподобия. К той же оценке приводит и метод моментов.

4. Задача оценивания параметра обсуждается в [47, с. 178]. Пример приведен в параграфе 10.12.

Литература: [47, с. 175—177].

13.12. СРЕДНЕЕ НОРМАЛЬНОЙ СОВОКУПНОСТИ

Обозначение и статистическая модель. См. параграф 12.13. Объем выборки равен n, среднее равно x, выборочная дисперсия — s^2.

Точечная оценка.

$$A = x. \qquad (13.12.1)$$

95 %-ные доверительные границы. Приравняем (12.13.1) к верхней и нижней 2,5 %-ным точкам t_{n-1}-распределения.

Комментарии. 1. Иногда дисперсия генеральной совокупности σ^2 известна. Тогда вместо t-распределения необходимо использовать стандартное нормальное распределение, при этом нет необходимости вычислять значение s^2. В формуле (12.13.1) вместо s подставляется стандартное отклонение совокупности a.

2. Метод устойчив при умеренных отклонениях от нормальности.

Пример 13.12.1. Из нормальной генеральной совокупности с дисперсией 1 и неизвестным средним произведена выборка объема 9. Выборочное среднее равно 4,11, которое и является оценкой μ. Найдем 95 %-ный доверительный интервал для μ.

В соответствии с п. 1 комментариев приравняем (12.13.1) к верхней и нижней 2,5 %-ным точкам стандартного нормального распределения, заменяя s на a:

$$\sqrt{9}\,(4{,}11-\mu)/1 = \pm 1{,}96.$$

Решая уравнение относительно μ, находим, что 95 %-ный доверительный интервал есть $4{,}11 \pm 0{,}65$.

Литература: [2, с. 43—51; русский перевод с. 51—56], [3, с. 102—104], [45, с. 138], [66, с. 160—161], [70, с. 251—253], [81, с. 151—156], [93, с. 156—166], [102, с. 203—209; русский перевод с. 221—227].

13.13. РАЗНОСТЬ ДВУХ СРЕДНИХ — СЛУЧАЙ ОДИНАКОВЫХ ДИСПЕРСИЙ

Обозначение и статистическая модель. См. параграф 12.15. Обозначим $\delta = \mu_1 - \mu_2$.

Точечная оценка.

$$\hat{\delta} = \bar{x}_{.1} - x_{.2}. \qquad (13.13.1)$$

95 %-ные доверительные границы. Приравняем (12.15.1) к верхней и нижней 2,5 %-ным точкам t-распределения с $n_1 + n_2 - 2$ степенями свободы.

Комментарии. 1. Иногда общая дисперсия σ^2 известна. В этом случае вместо таблиц t-распределения воспользуемся таблицами стандартного нормального распределения, а вместо s_1^2 и s_2^2 — общей дисперсией σ^2.

2. Описанный метод устойчив при умеренных отклонениях от нормальности, а также при равенстве дисперсий при условии, что n_1 и n_2 приблизительно равны (см. также параграф 13.14).

3. В случае парных наблюдений этот метод применять нельзя (см. параграф 13.17).

Пример 13.13.1. Два сорта пшеницы сравниваются по урожайности. Сорт A — обычный сорт пшеницы, сорт B — новый гибрид. Каждым сортом пшеницы было засеяно по 25 акров; урожай получен при одинаковых условиях. Средний урожай пшеницы сорта A с одного акра составил 32,0 бушеля с дисперсией 5,9, средний урожай сорта B — 36,2 бушеля с дисперсией 11,2. Найдем оценку максимального правдоподобия разности средних урожаев двух сортов пшеницы и 95 %-ный доверительный интервал для этой разности.

Оценка по методу максимального правдоподобия $\hat{\delta} = \bar{x}_B - x_A = 4{,}2$. Для нахождения 95 %-ного доверительного интервала решим уравнения

$$\frac{[(36{,}2-32{,}0)-\delta]\sqrt{48}}{[(2/25)\,(24\times 5{,}9+24\times 11{,}2)]^{\frac{1}{2}}} = \pm 1{,}96.$$

Доверительными границами будут 2,58 и 5,82 бушеля.

Литература: [3, с. 129—133], [46, с. 264], [102, с. 212—215].

13.14. РАЗНОСТЬ ДВУХ СРЕДНИХ — СЛУЧАЙ РАЗНЫХ ДИСПЕРСИЙ

Обозначение и статистическая модель. См. параграф 12.16. Обозначим $\delta = \mu_1 - \mu_2$.

Точечная оценка.

$$\hat{\delta} = \bar{x}_{.1} - x_{.2}. \qquad (13.14.1)$$

95 %-ные доверительные границы. Приравниваем (12.16.2) к верхней и нижней 2,5 %-ным точкам t_ν-распределения (v определяется выражением (12.16.1)).

Комментарии. 1. Доверительный интервал является приближенным. Если нет доказательств того, что дисперсии различаются, то более предпочтительно применение точного метода из параграфа 13.13.

2. Метод устойчив при умеренных отклонениях от нормальности.

3. Если n_1+n_2 больше 20, то можно избежать вычисления v. При этом вместо t-распределения необходимо воспользоваться таблицами нормального распределения.

4. Этот метод применять нельзя, если наблюдения образуют пары (см. параграф 13.17).

Пример 13.14.1. Сравнивается урожайность двух сортов пшеницы. Данные эксперимента содержатся в примере 13.13 1. Найдем оценку разности урожаев двух сортов пшеницы по методу максимального правдоподобия и построим для нее 95 %-ный доверительный интервал.

Как следует из предыдущего примера, оценка максимального правдоподобия равна $\hat{\delta} = 4{,}2$. Для нахождения 95 %-ных границ решим уравнения

$$\frac{(36{,}2-32{,}0)-\delta}{(5{,}9/25+11{,}2/25)^{1/2}} = \pm 1{,}96,$$

которые приведут нас к интервалу (2,58; 5,82).

Литература: [2, с. 43—51; русский перевод с. 71—74], [3, с. 129—133], [7, с. 299—303; русский перевод с. 111—112], [46, с. 262—265].

13.15. ЛИНЕЙНАЯ КОМБИНАЦИЯ СРЕДНИХ * СОВОКУПНОСТЕЙ С ОБЩЕЙ ДИСПЕРСИЕЙ

Обозначение и статистическая модель. См. параграф 12.18. Средние генеральных совокупностей $\mu_1, \mu_2, \ldots, \mu_k$ неизвестны. Мы хотим оценить $L = c_1\mu_1+\ldots+c_k\mu_k$, где $\{c_i\}$ — константы, которые могут быть отрицательными, положительными или равными нулю. Среднее и дисперсия выборки i есть $\bar{x}_{\cdot i}$ и s_i^2 соответственно; $n = n_1+n_2+\ldots+n_k$.

Точечная оценка.

$$\hat{L} = c_1\bar{x}_{\cdot 1}+\ldots+c_k\bar{x}_{\cdot k}. \qquad (13.15.1)$$

95 %-ные доверительные **границы.** а) *Имеется одно множество значений* $\{c_i\}$, *выбранное до начала эксперимента.* Вычислим

$$s^2 = \{(n_1-1)s_1^2+\ldots+(n_k-1)s_k^2\}/(n-k). \qquad (13.15.2)$$

Приравняем

$$F = (\hat{L}-L)^2/\{(c_1^2/n_1+\ldots+c_k^2/n_k)s^2\} \qquad (13.15.3)$$

к верхней 5 %-ной точке F-распределения с 1 и $n-k$ степенями свободы.

б) *Имеется несколько различных множеств коэффициентов* $\{c_i\}$ *или они выбираются после изучения имеющихся выборок.* Формулу (13.15.3) в этом случае применять нельзя. Как правило, задача заключается в сравнении средних, в этом случае сумма коэффициентов c_i равна нулю и можно использовать формулу (13.15.4). Вычислим s^2 по формуле (13.15.2) и приравняем

$$S = \{1/(k-1)\}(\hat{L}-L)^2/\{(c_1^2/n_1+\ldots+c_k^2/n_k)s^2\} \qquad (13.15.4)$$

к верхней 5 %-ной точке F-распределения с $k-1$ и $n-k$ степенями свободы.

Комментарии. 1. Напомним, что, как следует из примера 9.4.2, случайная величина, распределенная по F-распределению с 1 и $n-k$ степенями свободы, является квадратом случайной величины, распределенной по t-распределению с $n-k$ степенями свободы. Доверительные границы, получаемые на основе процедуры а), можно определить также приравниванием квадратного корня из величины (13.15.3) к верхней и нижней 2,5 %-ным точкам t-распределения с $n-k$ степенями свободы. Процедуру а) можно, таким образом, назвать t-методом. Метод из параграфа 13.13 представляет собой частный случай.

2. Необходимо предупредить читателя, что вычисленные по тем же данным t-интервалы не будут иметь 95 %-ный уровень доверия; аналогично если коэффициенты $\{c_i\}$ выбираются после эксперимента, применение t-метода неправомерно. В начале параграфа 12.21 объясняется, почему это происходит. Процедура б) основана на многомерном методе сравнения Шеффе (см. параграф 12.21).

3. Устойчивость метода рассматривается в пунктах 2 и 3 комментариев из параграфа 12.18.

4. Приведенные выше методы могут применяться только в том случае, когда выборки имеют одинаковую дисперсию. Если дисперсии различаются, то вместо процедуры а) необходимо воспользоваться приближенным методом из параграфа 13.16.

5. Специальный случай, когда $c_1 = 1$, $c_2 = c_3 = \ldots = c_k = 0$, рассматривается Зароом в [105, с. 139].

6. Эти методы нельзя применять, если данные парные (см. параграф 13.18).

7. При одномерном анализе дисперсий s^2 равно остаточной сумме квадратов, а вычисление s_1^2, s_2^2 ..... s_k^2 не обязательно.

Пример 13.15.1. При изготовлении машины собираются в узел три детали. Две детали — одного и того же типа A, третья —

типа *B*. Необходимо оценить общую длину трех деталей в сборе. Выборочные значения равны:

$$n_A = 10; \quad x_A = 3{,}02; \quad s_A^2 = 0{,}0004;$$
$$n_B = 15; \quad x_B = 10{,}04; \quad s_B^2 = 0{,}0006.$$

Найдем оценку максимального правдоподобия средней длины трех деталей в сборе и ее 95 %-ный доверительный интервал. Оценка максимального правдоподобия следующая:

$$Z = 2 \text{ X } 3{,}02 + 1 \text{ X } 10{,}04 = 16{,}08.$$

Оценка дисперсии: $s^2 = (9\text{X}0{,}0004 + 14\text{x}0{,}0006)/23 = 0{,}00052$. Для нахождения 95 %-ного доверительного интервала для L воспользуемся процедурой а), для чего решим уравнение

$$(16{,}08 - L)^2/\{(2^2/10 + 1^2/15)\ (0{,}000\ 52)\} = 4{,}28,$$

которое приводит к интервалу $16{,}08 \pm 0{,}03$.

Литература: [7, с. 474], [46, с. 294], [70, с. 374], [88, с. 30; русский перевод с. 31], [105, с. 151—162].

13.16. ЛИНЕЙНАЯ КОМБИНАЦИЯ СРЕДНИХ k СОВОКУПНОСТЕЙ С РАЗНЫМИ ДИСПЕРСИЯМИ

Обозначение и статистическая модель. См. параграф 13.15. Единственное множество констант $\{c_i\}$ выбрано априори до начала эксперимента.

Точечная оценка.

$$\hat{L} = c_1\bar{x}_{\cdot 1} + \ldots + c_k\bar{x}_{\cdot k}. \qquad (13.16.1)$$

95 %-ные доверительные интервалы. Приравняем

$$F = (\hat{L} - L)^2/(c_1^2 s_1^2/n_1 + \ldots + c_k^2 s_k^2/n_k) \qquad (13.16.2)$$

к верхней 5 %-ной точке F-распределения с 1 и v степенями свободы, где параметр v определяется выражением

$$\nu = \frac{\left(\frac{c_1^2 s_1^2}{n_1} + \ldots + \frac{c_k^2 s_k^2}{n_k}\right)^2}{\left(\frac{c_1^2 s_1^2}{m}\right)^2 \frac{1}{n_1 - 1} + \ldots + \left(\frac{c_k^2 s_k^2}{n_k}\right)^2 \frac{1}{n_k - 1}} - 2. \qquad (13.16.3)$$

Комментарии. 1. Этот приближенный метод представляет собой аналог процедуры а) из предыдущего параграфа для случая, когда выборки имеют разные дисперсии. При равных дисперсиях более предпочтительна процедура а).

2. Метод устойчив при умеренных отклонениях от нормальности.

3. Напомним, что, как следует из примера 9.4.2, случайная величина, распределенная по F-распределению с 1 и v степенями свободы, является квадратом случайной величины, распределенной по t-распределению с v степенями свободы. Те же 95 %-ные доверительные границы могут быть получены приравниванием квадратного корня из (13.16.2) к верхней и нижней 2,5 %-ным точкам t-распределения с v степенями свободы. Метод из параграфа 13.14 представляет собой особый случай.

4. При достаточно больших n (например, при $n > 25$) можно избежать вычисления v и вместо F-распределения с 1 и v степенями свободы применить распределение χ^2.

5. Этот метод непригоден при зависимых выборках (см. параграф 13.18).

Пример 13.16.1. При изготовлении машины собираются в узел три детали. Две детали — одного и того же типа *A*, третья — типа *B*. Необходимо найти оценку общей длины трех деталей в сборе. Получены следующие выборочные данные:

$$n_A = 10; \quad \bar{x}_A = 3{,}04; \quad s_A^2 = 0{,}0004;$$
$$n_B = 15; \quad x_B = 10{,}15; \quad s_B^2 = 0{,}0095.$$

Найдем оценку максимального правдоподобия:

$$\hat{L} = 2\text{X}3{,}04 + 1\text{ X}10{,}15 = 16{,}23.$$

Общий объем выборки $n = 25$, поэтому вместо $F_{1,\nu}$-распределения можно взять χ_1^2-распределение. Для получения 95 %-ных доверительных границ для L решим уравнение

$$(16{,}23 - L)^2 \Big/ \left(\frac{2^2 \times 0{,}0004}{10} + \frac{1^2 \times 0{,}0095}{15}\right) = 3{,}84.$$

Доверительными границами будут 16,17 и 16,29.

Литература: [7, с. 299—303; русский перевод с. 290—293].

13.17. СРЕДНЕЕ РАЗНОСТИ ПАРНЫХ НАБЛЮДЕНИЙ

Обозначение и статистическая модель. См. параграф 12.22.

Точечная оценка.

$$\hat{\mu} = \text{rf}. \qquad (13.17.1)$$

95 %-ные доверительные границы. Приравняем (12.22.1) к верхней п нижней 2,5 %-ным точкам t_{n-1}-распределения.

Комментарии. 1. Метод устойчив при умеренных отклонениях от нормальности.

2. Метод неприменим для непарных наблюдений.

Пример 13.17.1. Рассмотрим эксперимент с пшеницей из примера 12.21.1. Объем выборки равен 10, а средняя разность урожаев двух сортов пшеницы составляет $\bar{d} = 0{,}57$, что является оценкой максимального правдоподобия разности урожаев. Выборочная дисперсия $s^2 = 0{,}4312$. Вычислим 95 %-ные доверительные границы разности урожаев.

Для построения доверительных границ необходимо решить уравнение

$$\sqrt{10}\,(0{,}57 - \mu)/\sqrt{0{,}4312} = \pm 2{,}262.$$

Этими границами являются $0{,}57 \pm 0{,}47$.

Литература: [3, с. 129—133], [105, с. 121—124].

13.18. СРЕДНИЕ НЕСКОЛЬКИХ ЗАВИСИМЫХ (ПАРНЫХ) СОВОКУПНОСТЕЙ

Обозначение и статистическая модель. См. параграф 12.25. Эффекты по строке $\alpha_1, \ldots, \alpha_r$ неизвестны, необходимо оценить $L = c_1\alpha_1 + \ldots + c_r\alpha_r$, где константы c_i в сумме равны нулю. Остаточную сумму квадратов отклонений в табл. 12.25.1 дисперсионного анализа обозначим s^2. Последовательность средних по строке $x_{1.} = T_{1.}/c, \ldots, x_{r.} = T_{r.}/c$.

Точечная оценка.

$$\hat{L} = c_1\bar{x}_{1.} + \ldots + c_r\bar{x}_{r.}. \qquad (13.18.1)$$

95 %-ные доверительные границы. а) *Имеется одно множество значений коэффициентов, выбранное до начала эксперимента.* Приравняем

$$F = c(\hat{L} - L)^2/\{(c_1^2 + \ldots + c_r^2)\,s^2\} \qquad (13.18.2)$$

к верхней 5 %-ной точке F-распределения с 1 и $(r-1)(c-1)$ степенями свободы.

б) *Имеется несколько различных множеств коэффициентов или они выбираются после анализа данных эксперимента.* Приравняем

$$5 = \{c/(r-1)\}\,(\hat{L} - L)^2/\{(c_1^2 + \ldots + c_r^2)\,s^2\} \qquad (13.18.3)$$

к верхней 5 %-ной точке F-распределения с $r-1$ и $(r-1)(c-1)$ степенями свободы.

Комментарии. 1. Напомним, что, как следует из примера 9.4.2, случайная величина $F_{1,n}$ является квадратом величины t_n. Доверительные границы, получаемые на основе процедуры а) можно также построить приравниванием квадратного корня из (13.18.2) к верхней и нижней 2,5 %-ным точкам t-распределения с $(r-1) \times (c-1)$ степенями свободы. Таким образом, процедуру а) можно назвать t-методом. Метод из параграфа 13.17 представляет собой особый случай.

2. Использование нескольких 95 %-ных t-интервалов, вычисленных на основе одних и тех же данных, некорректно, ошибочно также применение t-метода в том случае, когда коэффициенты $\{c_i\}$ выбираются после анализа данных. В начале параграфа 12.21 объясняются причины этого факта. Процедура б) основана на многомерном методе сравнения Шеффе (см. параграф 12.27).

3. Устойчивость метода обсуждается в п. 2 комментариев из параграфа 12.25.

4. Для применения процедуры к эффектам по столбцам необходимо r и c поменять местами и использовать средние по столбцам $\{\bar{x}_{.j}\}$.

Пример 13.18.1. В табл. 12.25.3 показано содержание фосфора (мг/100 г) в сердце, легких, печени и почках животных трех пород. Найдем 95 %-ный доверительный интервал для разности между содержанием фосфора в сумме в сердце и печени, в легких и почках.

В этом примере обратимся к эффектам по столбцам, положив $c_1 = C3 = 1$, $C2 = c_4 = -1$. Средние по столбцам равны 85,4; 103,5; 206,3; 182,3. Оценка максимального правдоподобия, таким образом, будет следующей:

$$Z = 85{,}4 - 103{,}5 + 206{,}3 - 182{,}3 = 5{,}9.$$

Данные приведенного примера использованы в параграфе 12.25, а коэффициенты были выбраны в результате предварительного анализа этих данных, поэтому необходимо применить процедуру б). В соответствии с таблицей дисперсионного анализа 12.25.4 остаточная сумма квадратов отклонений $s^2 = 5{,}061$. Для нахождения 95 %-ных доверительных границ для L следует решить уравнение

$$[3/(4-1)]\,(5{,}9 - L)^2/(4 \times 5{,}061) = 4{,}76.$$

Искомые границы равны $5{,}9 \pm 9{,}8$.

Литература: [39, с. 79], [88, с. 30], [105, с. 151 — 162].

13.19. ДИСПЕРСИЯ НОРМАЛЬНОЙ СОВОКУПНОСТИ С НЕИЗВЕСТНЫМ СРЕДНИМ

Обозначение и статистическая модель. Исследуемая генеральная совокупность является нормальной с неизвестным средним и неизвестной дисперсией σ^2. Выборка $x_1, \ldots, x_n$ имеет объем n и выборочную дисперсию s^2.

Точечная оценка.

$$\hat{\sigma}^2 = s^2. \qquad (13.19.1)$$

95 %-ные доверительные границы. Приравняем (12.28.1) к верхней и нижней 2,5 %-ным точкам распределения χ^2_{n-1}.

Комментарии. 1. Если среднее совокупности известно, то необходимо воспользоваться методом из параграфа 13.20.

2. Метод не очень устойчив при отклонениях от нормальности.

Пример 13.19.1. В параграфе 12.28 приведен пример с ручной гранатой. Объем выборки равен 11, а выборочная дисперсия времени срабатывания запала равна 0,017. Таким образом, несмещен-

ная оценка σ^2 равна 0,017; 95 %-ные доверительные границы находятся из решения уравнений

$$\frac{10 \times 0{,}017}{\sigma^2} = 20{,}48 \text{ и } \frac{10 \times 0{,}017}{\sigma^2} = 3{,}25.$$

Эти границы равны 0,0083 и 0,0523.

Литература: [3, с. 104—106, 108—109], [46, с. 254—257] [70, с. 254—255], [105, с. 97].

13.20. ДИСПЕРСИЯ НОРМАЛЬНОЙ СОВОКУПНОСТИ С ИЗВЕСТНЫМ СРЕДНИМ

Обозначение и статистическая модель. Исследуемая генеральная совокупность является нормальной с известным средним μ, но неизвестной дисперсией σ^2. Выборка $x_1, \ldots, x_n$ имеет объем n и выборочное среднее $\bar{x}$.

Точечная оценка.

$$\hat{\sigma}^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu(2\bar{x} \text{ -ц}). \qquad (13.20.1)$$

95 %-ные доверительные границы. Приравняем (13.20.2) к верхней и нижней 2,5 %-ным точкам распределения χ^2_n:

$$\chi^2 = nS^2/\sigma^2. \qquad (13.20.2)$$

Комментарии. 1. При неизвестном среднем μ совокупности необходимо применять метод из параграфа 13.19.

2. Метод не очень устойчив при отклонениях от нормальности.

Пример 13.20.1. Определяется квалификация лаборанта. Ему дается латунная деталь, длина которой известна и равна 6,764 см; при этом его просят произвести 10 измерений микрометром. Результаты измерений следующие: 6,772; 6,763; 6,759; 6,768; 6,771; 6,764; 6,756; 6,762; 6,761; 6,770. Найдем 95 %-ный доверительный интервал для дисперсии десяти измерений лаборанта.

С помощью (13.20.1) найдем оценку максимального правдоподобия, которая равна 268×10^{-7}. Для нахождения 95 %-ных доверительных границ решим уравнения

$$\frac{10 \times 268 \times 10^{-7}}{\sigma^2} = 20.48 \text{ и } \frac{10 \times 268 \times 10^{-7}}{\sigma^2} = 3{,}25.$$

Этими границами являются 131×10^{-7} и 825×10^{-7}.

Комментарий. При предположении относительно смещения результатов измерения, по-видимому, желательнее воспользоваться методом из параграфа 13.12 для оценки среднего измерений и методом из параграфа 13.19 для оценки дисперсии.

13.21. ОБЩАЯ ДИСПЕРСИЯ НЕСКОЛЬКИХ СОВОКУПНОСТЕЙ

Обозначение и статистическая модель.

Имеется k независимых нормально распределенных совокупностей с неизвестными средними и неизвестной общей дисперсией σ^2.

Объем выборки из i-й совокупности равен n_i, выборочная дисперсия равна s^2. Значение s^2 находится по формуле **(13.15.2).**

Точечная оценка.

$$\hat{\sigma}^2 = s^2. \qquad (13.21.1)$$

95 %-ные доверительные границы. Приравняем (13.21.2) к верхней и нижней 2,5 %-ным точкам распределения χ^2_{n-k}

$$\chi^2 = \frac{(n-k)\,s^2}{\sigma^2}. \qquad (13.21.2)$$

Комментарии. 1. При однофакторном анализе дисперсионной таблицы s^2 равно остаточной сумме квадратов, вычисление $s_1^2, \ldots, s_k^2$ не обязательно (см. параграф 12.18).

2. Метод не очень устойчив при отклонениях от нормальности.

Пример 13.21.1. При изготовлении машины применяются детали двух различных типов *А* к *В*. Нас интересует длина этих деталей, при этом известно, что они имеют одинаковую дисперсию. Были получены следующие данные: тип *А:* объем выборки равен 10, среднее 3,02, дисперсия 0,0004; тип *В:* объем выборки равен 15, среднее 10,04, дисперсия 0,0006.

Оценим общую дисперсию двух деталей. Точечная оценка этой неличины следующая:

$$s^2 = (9 \text{ x } 0{,}0004 + 14 \text{ X } 0{,}0006)/23 = 0{,}00052.$$

Для определения доверительных границ решим уравнения

$$\frac{23 \times 0{,}00052}{\sigma^2} = 38{,}08 \text{ и } \frac{23 \times 0{,}00052}{\sigma^2} = 11{,}69.$$

Доверительные границы равны 0,00031 и 0,00102.

Литература: [3, с. 111, 159],

13.22. ПАРНЫЕ НАБЛЮДЕНИЯ. ДИСПЕРСИЯ РАЗНОСТИ

Обозначение и статистическая модель. См. параграф 12.22.

Точечная оценка.

$$\hat{\sigma}^2 = s_d^2. \qquad (13.22.1)$$

95 %-ные доверительные границы. Приравняем

$$\chi^2 = (n-1)\,s_d^2/\sigma^2 \qquad (13.22.2)$$

к верхней и нижней 2,5 %-ным точкам распределения χ^2_{n-1}.

Комментарии. 1. Случай нескольких зависимых выборок рассматривается в параграфе 13.23.

2. Метод не очень устойчив при отклонениях от нормальности.

Пример 13.22.1. Оценим дисперсию разности урожаев пшеницы двух сортов Л и В из примера 12.22.1.

Точечная оценка $s^2_d = 0{,}4312$. Доверительные границы получаются из решения уравнений

$$9 \times 0{,}4312/\sigma^2 = 19{,}02 \text{ и } 9 \times 0{,}4312/\sigma^2 = 2{,}70.$$

Эти границы равны 0,204 и 1,437.

13.23. ЗАВИСИМЫЕ НАБЛЮДЕНИЯ. ДИСПЕРСИЯ ОШИБКИ e_{ij} ПРИ ДВУХФАКТОРНОМ ДИСПЕРСИОННОМ АНАЛИЗЕ

Обозначение и статистическая модель. См. параграф 12.25. Остаточный средний квадрат из таблицы дисперсионного анализа 12.25.1 обозначим через s^2.

Точечная оценка.

$$\hat{\sigma}^2 = s^2. \tag{13.23.1}$$

95 %-ные доверительные границы. Приравняем

$$\chi^2 = (r-1)(c-1)\, s^2/\sigma^2 \tag{13.23.2}$$

к верхней и нижней 2,5 %-ным точкам распределения χ^2 с $(r-1) \times (c-1)$ степенями свободы.

Комментарии. 1. В случае двух зависимых выборок можно применять метод из параграфа 13.22. Два подхода математически эквивалентны.

2. Метод не очень устойчив при отклонениях от нормальности.

Пример 13.23.1. Исследуется содержание фосфора (мг/100 г) в каждом из четырех органов животных трех пород. Результаты исследования приведены в табл. 12.25.3. Оценим дисперсию ошибки.

В соответствии с таблицей дисперсионного анализа 12.25.4 остаточный средний квадрат равен 5,061, что в свою очередь представляет собой оценку σ^2. Для нахождения 95 %-ных доверительных границ решаем уравнения

$$6 \times 5{,}061/\sigma^2 = 14{,}45 \text{ и } 6 \times 5{,}061/\sigma^2 = 1{,}24,$$

которые приводят к значениям 2,1 и 24,5.

Литература: [3, с. 168].

13.24. ОТНОШЕНИЕ ДВУХ ДИСПЕРСИИ

Обозначение и статистическая модель. Имеются две независимые нормальные совокупности с неизвестными средними. Неизвестны также дисперсии совокупностей σ^2_1 и σ^2_2. Необходимо оценить отношение $\gamma = \sigma^2_1/\sigma^2_2$. Из соответствующих совокупностей извлечены выборки объемов n_1 и n_2, при этом выборочные дисперсии соответственно равны s^2_1 и s^2_2.

Точечная оценка.

$$\hat{\gamma} = \frac{n_2 - 3}{n_2 - 1} \frac{s^2_1}{s^2_2}. \tag{13.24.1}$$

95 %-ные доверительные границы. Приравняем (12.29.1) к верхней и нижней 2,5 %-ным точкам F-распределения с $n_1 - 1$ и $n_2 - 1$ степенями свободы.

Комментарии. 1. Нахождение нижних точек F-распределения показано в примере 9.4.3.

2. Метод не очень устойчив при отклонениях от нормальности.

Пример 13.24.1. Дисперсии двух независимых выборок равны 11,2 и 4,7. Объемы выборок соответственно равны 17 и 51. Оценим отношение дисперсии первой выборки к дисперсии второй.

Точечная оценка $\hat{\gamma} = (48/50)\,(11{,}2/4{,}7) = 2{,}29$. Для нахождения 95 %-ных доверительных границ для отношения $\gamma = \sigma^2_1/\sigma^2_2$ решим уравнения

$$(11{,}2/\sigma^2_1)/(4{,}7/\sigma^2_2) = 2{,}1 \text{ и } (4{,}7/\sigma^2_2)/(11{,}2/\sigma^2_1) = 2{,}48.$$

Границами являются 1,13 и 5,91.

Литература: [3, с. 145—150], [105, с. 101—103].

13.25. ПАРАМЕТРЫ ЛОГАРИФМИЧЕСКИ-НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Параметрами логарифмически-нормального распределения являются 0, μ и σ^2 (см. параграф 9.5). При известном значении параметра 6 (часто Э можно положить равным нулю) оценивание μ и σ^2 можно производить непосредственно. Действительно, полагая $z_i = \ln(x_i - 6)$ и используя $\{zi\}$ вместо прежнего набора данных $\{x_i\}$, сведем задачу к оцениванию среднего и дисперсии нормального распределения (см. параграфы 13.12 и 13.19). Точечными оценками μ и σ^2 соответственно являются выборочное среднее z и выборочная дисперсия s^2 данных $\{z_i\}$. Доверительные интервалы для μ и σ^2 могут быть найдены с помощью методов, описанных в параграфах 13.12 и 13.19, и преобразованных данных $\{z_i\}$.

При неизвестном 0 задача оценивания усложняется. Параметр I) является пороговым, ниже этого значения функция распределения равна нулю, выше — положительна. Оценки 0, как правило, весьма не точны. Однако оценивание частных параметров обычно не столь важно, как точное оценивание кумулятивной функции распределения. Оказывается, что последние не очень чувствительны к оценкам 0 [47, с. 122]. Представляется, что в качестве оценки следует выбирать такое значение 0, которое несколько меньше минимального наблюдения x, после чего μ и σ^2 оценить обычным способом. При больших объемах выборки оценка 0 должна быть очень близкой к наименьшему наблюдению.

Литература: [47, с. 119—127].

13.26. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ρ

Обозначение и статистическая модель. См. параграф 12.31. Выборочный коэффициент корреляции равен r, а объем выборки равен n.

Точечная оценка.

$$\hat{\rho} = r. \qquad (13.26.1)$$

95 %-ные доверительные границы. Для выражения z через r воспользуемся преобразованием Фишера (см. параграф 12.31). Приравняем

$$T = \sqrt{n-3}\,(z-\zeta) \qquad (13.26.2)$$

к верхней и нижней 2,5 %-ным точкам стандартного нормального распределения, откуда получим доверительные границы для ζ (преобразование р). Затем применим обратное преобразование Фишера к ζ для получения доверительных границ для р. При этом очень полезной будет табл. 12.31.1.

Комментарии. 1. Описанная процедура приводит к приближенным доверительным границам *.

2. Устойчивость метода обсуждается в п. 3 комментариев из параграфа 12.31.

Пример 13.26.1. В табл. 12.31.2 показаны оценки за ответы пятнадцати студентов на два экзаменационных вопроса. Точечная оценка корреляции между оценками за ответы на вопрос 1 и на вопрос 2 определяется с помощью выборочного коэффициента корреляции r, который в данном случае равен 0,764. Соответствующее значение z равно 1,006. Для нахождения 95 %-ных доверительных границ для ζ решим уравнения

$$\sqrt{12}\,(1{,}006-\zeta) = \pm 1{,}96.$$

Границами для £ являются 0,440 и 1,572, поэтому границы для р соответственно равны 0,413 и 0,917.

13.27. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ ДЛЯ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Обозначение и статистическая модель. Дана совокупность с неизвестной непрерывной функцией распределения $F(x)$. Выборка n независимых наблюдений $x_1, \ldots, x_n$ ранжирована, т. е. $x_1 < x_2 < \ldots < x_n$.

95 %-ные **доверительные границы.** Для построения 95 %-ных доверительных границ для $F(x)$ в точке Xj приравняем

$$Z = |F(x_j) - j/n| \qquad (13.27.1)$$

к верхней 5 %-ной точке распределения Колмогорова—Смирнова (см. 12.21.1).

Комментарии. 1. Эта процедура дает 95 %-ные доверительные границы для $F(x)$ в каждой из n точек выборки.

* В том смысле, что построенный доверительный интервал накрывает истинное значение р с вероятностью, близкой к 0.95, но не равной этому значению точно.— *Примеч. пер.*

2. Избегайте применять описанный метод для данных из дискретных распределений или сгруппированных непрерывных данных. Доверительный интервал, полученный в этом случае, будет шире истинного 95 %-ного доверительного интервала. Детали обсуждаются в параграфе 12.12.

Пример 13.27.1. В табл. 12.15.1 приводятся данные об уровне сывороточного холестерола у семи самцов черепахи. Эти же данные представлены в первом столбце табл. 13.27.1. Для построения 95 %-ных границ для функции распределения уровня сывороточного холестерола в точке $x = 222{,}5$ прибавим 0,483 к $^3/_7$ и вычтем 0,483 из $^3/_7$.

Таблица **13.27.1. Доверительные границы для функции распределения уровня сывороточного холестерола у семи самцов черепахи**

Уровень сывороточного холестерола (x_j)	Число наблюдений, меньших или равных x_j (Л	95%-ные доверительные границы для $F(x)$	
		нижняя	верхняя
218,6	1	0,000	0,626
220,1	2	0,000	0,769
222,5	3	0,000	0,912
224,1	4	0,088	1,000
226,5	5	0,231	1,000
228,8	6	0,374	1,000
229,6	7	0,517	1,000

Верхняя граница, таким образом, равна 0,912. Вычисленная нижняя граница отрицательна. Однако мы знаем, что значение функции распределения лежит на интервале (О, 1), поэтому в качестве нижней границы берем нуль. Доверительные границы в остальных шести точках вычисляются аналогично.

Литература: [54, с. 457—458].

13.28. УПРАЖНЕНИЯ

1. В биномиальном эксперименте произведено 30 испытаний, из них 9 были успешными. В другом таком же эксперименте было произведено 40 испытаний, при этом 18 из них оказались успешными. Найдите 95 %-ный доверительный интервал для разности между вероятностями успехов двух экспериментов.

2. Среднее пятнадцати независимых наблюдений из совокупности с пуассоновским распределением равно 0,6. Найдите 95 %-ные доверительные границы для параметра пуассоновского распределения K.

3. Найдите 95 %-ные доверительные границы для общей дисперсии трех совокупностей червей из примера 12.18.1.

4. Из трехпараметрического логарифмически-нормального распределения получена следующая выборка: 10,78; 6,79; 6,20; 3,74; 5,21; 5,93; 4,11; 6,97; 9,99; 18.43; 6.55; 11,36; 5,28; 9,99; 11,62; 6,72; 6,79; 4,84; 6,90; 8,24. Оцените параметры распределения.

5. Используя результаты из параграфа 12.32, предложите метод получения доверительных границ для разности двух коэффициентов корреляции.

14. НЕКОТОРЫЕ СПЕЦИАЛЬНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Вопросы, рассматриваемые в этой главе, можно разделить на три части: случайные числа (параграфы 14.1—14.3), преобразования данных (параграфы 14.4—14.7), цензурированные и усеченные распределения (параграфы 14.8—14.12). Тема «случайные числа» включает следующие вопросы: генерирование и использование случайных чисел на единичном интервале, преобразование случайных чисел для получения других распределений. В параграфах, посвященных преобразованиям данных, описываются преобразования арксинуса, квадратного корня, логарифмическое и обратные преобразования. В заключительной части главы рассматривается метод максимального правдоподобия для оценивания параметров цензурированного и усеченного пуассоновского и нормального распределений.

14.1. СЛУЧАЙНЫЕ ЧИСЛА

В статистическом анализе часто возникают задачи, которые требуют привлечения случайных чисел. При выборочных обследованиях такие числа могут использоваться для случайного отбора элементов из генеральной совокупности с целью оценивания ее параметров. В экспериментальной агрономии бывает необходимо на разных участках провести различные обработки (например, внести удобрения) случайным образом, для того чтобы при сравнении эффектов от обработок избежать смещения. Выбор участков основывается на случайных числах. Даже теоретики часто прибегают к случайным числам. Допустим, необходимо изучить распределение сложной случайной величины, которая определена в терминах других случайных величин с известными распределениями. С помощью случайных чисел будем генерировать случайные наблюдения с известным распределением, на основе которых затем вычислим случайные наблюдения сложной случайной переменной (см. пример 14.1.1). На основе этих выборочных значений теоретик-исследователь может построить неизвестную функцию распределения и ее характеристики (форму распределения, моменты и т. п.). Этот подход в 1908 г. применил У. С. Госсет (Стьюдент) для исследования t-распределения и распределения выборочного коэффициента корреляции [94], [95]. Термин *метод Монте-Карло* относится к разделу математики, связанному с применением случайных (рандомизированных) методов.

Много таблиц случайных чисел было опубликовано (например, [25, с. 134—139], [85], [105, с. 577—580]). Эти таблицы составлены из равновероятных десятичных цифр, объединенных в группы. Такие группы содержат, как правило, по пяти цифр, однако в некоторых таблицах число цифр в группах менее пяти. Подобные цифры, пары цифр или группы можно использовать в целях выбора. Например, для того чтобы из совокупности в 926 человек случайным образом отобрать 30, мы можем всех людей из этой совокупности пронумеровать от 1 до 926, а затем из таблицы случайных чисел выбрать 30 трехразрядных чисел. Если выбранное трехразрядное число не принадлежит интервалу от 1 до 926, то нужно выбрать три другие цифры.

Случайная дробь между 0 и 1 может быть получена выбором группы цифр (например, из 5 или из 10 цифр), причем запятая должна находиться перед выбранной группой. Обычно мы хотим, чтобы распределение такой случайной дроби было непрерывным, однако ясно, что дробь может принимать только значения, помноженные на 10^{-n}, где n — число цифр в дроби.

С таблицами случайных чисел необходимо обращаться очень осторожно. Случайность отбора цифр должна гарантироваться от одного выбора к другому. Один метод заключается в том, чтобы при обращении к таблице начинать с ее начала и фиксировать выбранные числа. В следующий раз можно начать с первого, неиспользованного ранее числа, таким образом постепенно продолжая процесс. В качестве другого метода можно предложить следующую процедуру. Открываем таблицу и не глядя отмечаем в ней пальцем место, которое соответствует некоторой группе чисел. Используя эти числа, выбираем страницу, строку, столбец и направление, в котором необходимо прочитать случайные числа.

Случайные числа генерируются также с помощью ЭВМ или программных калькуляторов. Эти числа, как правило, представляют собой дроби, заключенные между 0 и 1 и распределенные равномерно на всем единичном интервале. При этом мы обычно считаем распределение непрерывным, хотя нам известно, что получаемые с помощью ЭВМ числа пропорциональны некоторой малой величине, например 2^{-30}. Подобные числа лучше называть *псевдослучайными,* поскольку они получаются по определенному правилу. После того как первое число выбрано, можно построить далее всю их последовательность. Математические методы построения последовательностей случайных чисел применяются таким образом, что ни один статистический критерий не может обнаружить отсутствие случайности выбора. Большое достоинство последовательностей случайных чисел, построенных по заранее определенным правилам, заключается в том, что эти последовательности в целях проверки на случайность могут быть всегда воспроизведены. Первым методом генерирования псевдослучайных чисел был метод среднего квадрата: в качестве следующего случайного числа бралась цифра квадрата предыдущего случайного числа. Однако скоро выяснилось, что этот метод неудовлетворителен. Наиболее широко распространены *линейные сопряженные генераторы.* Эти генераторы оперируют целыми числами, а случайные переменные с распределением на единичном интервале затем получаются делением.

В линейном сопряженном генераторе первого порядка случайное целое на шаге $n-1$ (обозначим его как W_{n-1}) умножается на целую константу k_1, а произведение делится на другую целую кон-

станту p. Остаток от деления является целым числом, которое выбирается в качестве случайного целого на шаге $n(W_n)$. Данный процесс описывается уравнением

$$W_n = k_1 W_{n-1} \pmod{p}. \qquad (14.1.1)$$

Для получения случайной переменной, распределенной на единичном интервале, необходимо случайное целое W_n разделить на p. Константы p и k_1 следует выбирать с большой осторожностью. Дж. Г. Скеллам в некоторых неопубликованных работах рекомендует брать p = 999 563 и $k_1 = 470\,001$. При этом достигается длина полного цикла, равная $p - 1$.

Линейные сопряженные генераторы второго и третьего порядков задаются следующими уравнениями:

$$W_n = k_1 W_{n-1} + k_2 W_{n-2} \pmod{p}, \qquad (14.1.2)$$

$$W_n = k_1 W_{n-1} + k_2 W_{n-2} + k_3 W_{n-3} \pmod{p}. \qquad (14.1.3)$$

В каждом случае для получения случайной переменной с распределением на единичном интервале необходимо W_n разделить на p. Константы p, k_1, k_2, рекомендованные Дж. Г. Скелламом, приводятся в табл. 14.1.1.

Таблица 14.1.1. Целые константы, рекомендованные Дж. Г. Скелламом, при использовании линейных сопряженных генераторов. В каждом случае достигается полный цикл с длиной $p^s - 1$ (s — порядок генератора)

Порядок генератора (s)	Рекомендуемые константы			
	p	k_1	k_2	k_3
1	999 563	470 001	—	—
2	998917	366 528	508531	—
2	999 563	254 754	529 562	—
3	997 783	360 137	519815	616087
3	997 783	286 588	434 446	388 251

Пример 14.1.1. Число успехов в биномиальном эксперименте, в котором $n = 20$ и $p = 0{,}4$, является случайной величиной X, причем распределение X хорошо известно (см. параграф 10.1). Случайная переменная Y получена на основе выборки из биномиальной совокупности с параметрами $n = (1+X)^2$, $p = 1/(1+X)$, где X — случайная переменная, определенная выше. Распределение У довольно сложно, однако оно может быть получено эмпирически следующим образом:

1. Генерируем случайное число на единичном интервале.
2. Будем считать «успехом» ситуацию, когда случайное число примет значение, которое меньше или равно 0,4, в противном случае — «неудача».

3. Повторим шаги 1 и 2 двадцать раз. Число успехов в двадцати испытаниях представляет собой наблюдение за случайной переменной X. Вычислим $(1+X)^2$ и $1/(1+X)$.
4. Генерируем случайное число на единичном интервале.
5. «Успехом» считаем ситуацию, когда случайное число меньше или равно $1/(1+X)$, в противном случае — «неудача».
6. Повторим шаги 4 и 5 $(1+X)^2$ раз. Число успехов в $(1+X)^2$ испытаниях представляет собой наблюдение случайной переменной У.
7. Повторим шаги 1—6 большое число раз N и обозначим наблюдения распределения У через $Y_1, Y_2, \ldots, Y_N$. Оценку распределения У можно получить построением соответствующих частот полученных наблюдений.

Описанная процедура легко может быть выполнена на ЭВМ или на большом программируемом калькуляторе.

Литература: [26, с. 24—25], [40, с. 1—9, с. 25—31], [53, с. 213—226], [105, с. 16—17, 577—580].

14.2. СЛУЧАЙНЫЕ ЧИСЛА СО СТАНДАРТНЫМ НОРМАЛЬНЫМ РАСПРЕДЕЛЕНИЕМ

Нормальное распределение занимает центральное место в статистической теории, поэтому случайные числа со стандартным нормальным распределением часто бывают необходимы. Эти числа могут быть получены с помощью таблицы случайных чисел и таблицы нормального интеграла. Из таблицы случайных чисел находится случайная дробь U, заключенная между 0 и 1, затем по таблице нормального распределения отыскивается такое X, что $\Phi(X) = U$. Тогда X будет случайной величиной, распределенной по нормальному закону, с нулевым математическим ожиданием и единичной дисперсией.

Этот же метод можно реализовать на ЭВМ или программируемом калькуляторе. Однако хранить в памяти машины таблицу нормального интеграла и каждый раз вычислять обратный интеграл весьма нецелесообразно. Поэтому обычно применяются другие методы.

Один из таких методов основан на центральной предельной теореме [1]. При применении этого метода находится сумма N случайных дробей из единичного интервала. Эта сумма является случайной величиной со средним $N/2$ и дисперсией $N/12$. Вычтем из суммы $N/2$ и разность разделим на $\sqrt{N/12}$. В результате случайная переменная будет иметь нулевое математическое ожидание и единичную дисперсию; при достаточно больших N распределение этой переменной приблизительно нормально. Если необходима высокая скорость вычислений и решаемая задача мало зависит от хвостов нормального распределения, то N можно положить равным 12.

[1] См. параграф 9.1.

Элегантный метод был предложен Боксом и Мюллером [6]. На единичном интервале независимо друг от друга выбираются две случайные дроби U_1 и U_2. Вычислим

$$X_1 = (-2 \ln U_1)^{\frac{1}{2}} \cos (2\pi U_2), \qquad (14.2.1)$$

$$X_2 = (-2 \ln U_1)^{\frac{1}{2}} \sin (2\pi U_2). \qquad (14.2.2)$$

Тогда X_1 и X_2 — независимые случайные переменные со стандартным нормальным распределением. Метод абсолютно точен в математически безупречной ситуации, при которой U_2 и U_1 — независимые равномерно распределенные случайные величины (в частности, непрерывные). На практике распределение U_1 и U_2 дискретно. Если генератор случайных чисел для выработки U_1 и U_2 имеет небольшой модуль p (например, меньше 1 000 000), то хвосты распределений X_1 и X_2 будут построены недостаточно точно [72], однако для большинства задач это обстоятельство не имеет значения. При методе Бокса—Мюллера достигается меньшая скорость, чем при процедуре, основанной на центральной предельной теореме.

Для построения случайной переменной, распределенной по нормальному закону с математическим ожиданием μ и дисперсией σ^2, следует сначала построить случайную переменную со стандартным нормальным распределением. Затем эту переменную необходимо домножить на а и к результату прибавить μ.

Пример 14.2.1. Построим случайную переменную с нормальным распределением, имеющим математическое ожидание 10 и дисперсию 4. Обратимся к таблице случайных чисел, откуда случайным образом найдем случайную дробь, скажем 0,31088. Интерполируя таблицу нормального интеграла, получим $\Phi(-0{,}493) = 0{,}310\,88$. Таким образом, значение случайной переменной с нормальным распределением, имеющим математическое ожидание 10 и дисперсию 4, равно: $10 + 2X(-0{,}493) = 9{,}014$.

Пример 14.2.2. В примере 14.1.1 описан метод генерирования случайной переменной, распределенной по биномиальному закону с параметрами $n = 20$ и $p = 0{,}4$. Этот метод хорошо «работает» при небольших значениях n, с увеличением n скорость при его применении падает. Для больших n биномиальный закон может быть аппроксимирован нормальным (см. параграф 10.2). Сначала мы генерируем случайную переменную со стандартным нормальным распределением, домножаем ее на $\sqrt{npq}$, прибавляем np и результат округляем до ближайшего целого числа.

Пример 14.2.3. При достаточно больших K (например, при $Я \geqslant 10$) для построения случайной величины, распределенной по закону Пуассона с параметром K, можно также воспользоваться аппроксимацией нормальным законом. Для этого генерируем сначала случайную переменную со стандартным нормальным распределением, домножаем ее на $\sqrt{\lambda}$, прибавляем Я и результат округляем до ближайшего целого числа.

Пример 14.2.4. Для генерирования случайной переменной с распределением χ^2_n можно сначала сгенерировать n независимых случайных переменных со стандартным нормальным распределением, возвести их в квадрат и сложить (см. параграф 9.2).

Пример 14.2.5. Случайная переменная с распределением χ^2_m и случайная переменная с распределением χ^2_n могут быть получены генерированием $n+m$ независимых случайных переменных, имеющих стандартное нормальное распределение, после чего следует воспользоваться способом, предложенным в предыдущем примере. Случайная переменная с $F_{m,n}$-распределением может быть получена делением переменной с распределением χ^2_m на переменную с распределением χ^2_n и домножением результата на n/m (см. параграф 9.4).

Литература: [40, с. 39—40].

14.3. СЛУЧАЙНЫЕ ПЕРЕМЕННЫЕ С НЕПРЕРЫВНОЙ ФУНКЦИЕЙ РАСПРЕДЕЛЕНИЯ

В параграфе 14.2 был описан метод генерирования случайной переменной со стандартным нормальным распределением. Этот же метод можно применять для генерирования переменных с любой непрерывной функцией распределения. Для построения случайной переменной с непрерывной функцией распределения $F(x)$ необходимо сначала сгенерировать случайную дробь U, заключенную между 0 и 1, а затем вычислить $X = F^{-1}(U)$. Тогда X имеет функцию распределения, равную $F(x)$.

Этот метод обычно быстро приводит к результату, если $F(x)$ имеет аналитическую форму. Если это не так, т. е. мы располагаем лишь таблицей значений функции $F(x)$, то необходимо применение других методов.

Пример 14.3.1. Экспоненциальное распределение на отрезке $(0, \infty)$ задается функцией плотности вероятностей

$$f(x) = \lambda e^{-\lambda x} \qquad (14.3.1)$$

или кумулятивной функцией распределения

$$P(x) = 1 - e^{-\lambda x} \qquad (14.3.2)$$

Параметр Я является положительным, причем $F^{-1}(y) = -\{\ln(1-y)\}/\lambda$.

Для получения случайной переменной X с экспоненциальным распределением параметра Я мы генерируем случайную дробь на единичном интервале и вычисляем $X = -\{\ln(1-U)\}/\lambda$.

Литература: [48, с. 207—232].

14.4. ПРЕОБРАЗОВАНИЕ ДАННЫХ

В большинстве статистических задач предполагается, что дисперсия случайной переменной X и ее среднее никак не связаны между собой (например, в дисперсионном анализе, см. параграфы 12.18 и 12.25, и в регрессионном анализе, см. часть III). Однако часто возникают ситуации, когда подобные предположения не имеют места (например, среднее и дисперсия пуассоновского распределения совпадают). Тогда необходимо преобразовать случайную переменную в другую случайную переменную так, чтобы среднее и дисперсия не были связаны. Подобные преобразования для биномиального и пуассоновского распределений описаны в параграфах 14.5 и 14.6. Помимо того, что преобразования делают среднее и дисперсию независимыми, они часто приводят также к распределениям, которые ближе к нормальным, чем исходные.

Литература: [7, с. 144—146], [51, с. 54—56], [105, с. 182—183].

14.5. ПРЕОБРАЗОВАНИЕ АРКСИНУСА (ИЛИ УГЛОВОЕ ПРЕОБРАЗОВАНИЕ) ДЛЯ БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ

В эксперименте с биномиальным распределением проводится n испытаний, из них r оказались успешными, наблюдаемая доля успехов $P = r/n$ и соответствующая случайная переменная имеют математическое ожидание p. Дисперсия P, равная $p(1-p)/n$, *зависит как* от среднего p, так и от числа испытаний n. С другой стороны, дисперсия

$$Z = \arcsin\sqrt{P} \qquad (14.5.1)$$

почти не зависит от p. Действительно, если углы измерять в радианах, то дисперсия Z будет приближенно равна $1/4n$; если утлы измерять в градусах, то дисперсия Z будет приближенно равна $821/n$. Математическое ожидание Z почти не зависит от я и приближенно равно $\arcsin\sqrt{p}$.

Иногда используется другой вариант преобразования арксинуса:

$$Y = 2\sqrt{n}\arcsin\sqrt{P}. \qquad (14.5.2)$$

Дисперсия Y почти не зависит от n и p. Если углы измерять в радианах, то дисперсия Y будет приближенно равна 1; если углы измерять в градусах, то дисперсия будет приближенно равна 3283. Математическое ожидание Y зависит от двух параметров.

Приведенные преобразования хорошо «работают» на всем интервале, за исключением крайних значений, которые принимают переменные (т. е. около 0 %- и 100 %-ных точек). Преобразование

$$W = \arcsin\sqrt{\left(r+\frac{3}{8}\right)\Big/\left(n+\frac{3}{4}\right)} \qquad (14.5.3)$$

удовлетворительно на всем отрезке изменения доли (см. [55, с. 114; русский перевод с. 129—132]).

Помимо того, что преобразование арксинуса приводит к независимости дисперсии от среднего, оно дает также распределение, которое намного ближе к нормальному, чем исходное.

Все ЭВМ и почти все программируемые и непрограммируемые калькуляторы извлекают квадратный корень и находят арксинус. Поэтому описанные преобразования могут быть произведены достаточно быстро. Могут оказаться полезными и опубликованные таблицы (см., например, [105, с. 505—509]).

Пример 14.5.1. В ходе эксперимента было обследовано около 20 растений, при этом фиксировалось число растений, достигших за данное время определенного уровня развития. Этот эксперимент проводился для трех видов почвы с двумя видами удобрений. Результаты обследования показаны в табл. 14.5.1. Можно ли утверждать, что доли растений, достигших за данное время определенного уровня развития, для разных удобрений различны?

Таблица 14.5.1. Число растений, достигших за данное время определенного уровня развития при разных условиях

Вид почвы	Удобрение *A*			Удобрение *B*		
	число наблюдений n	число растений, достигших определенного уровня развития r	доля растений, достигших определенного уровня развития $P=r/n$	число наблюдений n	число растений, достигших определенного уровня развития r	доля растений, достигших определенного уровня развития $P=r/n$
1	15	9	0,600	20	11	0,550
2	18	10	0,556	19	14	0,737
3	20	10	0,500	20	13	0,650

Для решения этой задачи прежде всего напрашивается применение парного t-критерия из параграфа 12.22 и двухфакторный дисперсионный анализ из параграфа 12.25. В обоих случаях требуется постоянная дисперсия. Однако дисперсия биномиальной доли P в таблице двухфакторного анализа 14.5.1 зависит от неизвестного среднего клетки таблицы. Если к этим биномиальным долям применить преобразование арксинуса (14.5.1), то дисперсии станут приблизительно одинаковыми, поскольку объем каждой выборки близок к 20. Преобразованные переменные будут также приближенно нормальными. Мы можем применить двухфакторный дисперсионный анализ из параграфа 12.25 (или, что то же самое, парный t-критерий из параграфа 12.22) к преобразованным данным, представленным в табл. 14.5.2. Значение F при одной и двух степенях свободы равно 1,7 и не является значимым. Таким образом, нет доказательств того, что доли растений, достигших определенного уровня развития за данное время, для разных удобрений различны.

Замечание. В соответствии с изложенной теорией, общая дисперсия преобразованных данных должна быть приближенно равной 1/(4x20)= 0,0125. Средний квадрат ошибки в таблице дисперсионного анализа равен приближенно 0,0084. Эти значения можно считать достаточно согласованными.

Литература: [7, с. 144—146], [51, с. 54—56], [55, с. 114; русский перевод с. 139—142], [97], [105, с. 185—186].

14.6. ПРЕОБРАЗОВАНИЕ КВАДРАТНОГО КОРНЯ ДЛЯ ПУАССОНОВСКОГО РАСПРЕДЕЛЕНИЯ

Дисперсия случайной переменной X, распределенной по закону Пуассона, равна ее средней (см. параграф 10.6). Для получения случайной переменной с несвязанными средним и дисперсией может быть использовано *преобразование квадратного корня* (см. [1]):

$$Y = \sqrt{X + \frac{3}{8}}. \qquad (14.6.1)$$

Таблица 14.5.2. Результат применения преобразования арксинуса (14.5.1) к долям из табл. 14.5.1 (углы измерены в радианах)

Вид почвы	Удобрения		Всего
	A	*B*	
1	0,8861	0,8355	1,7216
2	0,8411	1,0321	1,8732
3	0,7854	0,9377	1,7231
Всего	2,5126	2,8053	5,3179

Если среднее X несколько больше 4, дробь 3/8 может быть опущена. Преобразование

$$Y = \sqrt{X} + \sqrt{X+1} \qquad (14.6.2)$$

предпочтительнее для случая, когда среднее пуассоновского распределения меньше 3 [55, с. 90].

Преобразование квадратного корня (14.6.1) в действительности необходимо использовать для получения случайной переменной, у которой среднее и дисперсия не связаны, в то время как дисперсия наблюдаемой случайной величины X пропорциональна средней (пуассоновское распределение представляет собой важный случай распределения с коэффициентом пропорциональности $k = 1$). В общем случае дисперсия Y приближенно равна $k/4$.

Преобразование квадратного корня также имеет тенденцию приводить к распределениям, которые ближе к нормальному, чем исходное.

Пример 14.6.1. В табл. 14.6.1 представлено 29 случайных наблюдений из распределения Пуассона со средним 9. В этой же таблице показаны данные, преобразованные по квадратному корню (с поправкой $^3/_8$). Выборочная дисперсия преобразованных данных равна 0,251.

Таблица 14.6.1. Применение преобразования квадратного корня к данным, распределенным по закону Пуассона. X_i обозначает i-е значение наблюдения в распределении Пуассона

X_i	$\sqrt{X_i+\frac{3}{8}}$	X_i	$\sqrt{X_i+\frac{3}{8}}$	X_i	$\sqrt{X_i+\frac{3}{8}}$
14	3,79	11	3,37	8	2,89
7	2,72	11	3,37	13	3,66
13	3,66	6	2,52	7	2,72
8	2,89	12	3,52	11	3,37
5	2,32	10	3,22	14	3,79
10	3,22	12	3,52	9	3,06
5	2,32	7	2,72	7	2,72
4	2,09	7	2,72	8	2,89
7	2,72	7	2,72	6	2,52
16	4,05	10	3,22		
			Всего	265	88,30

Необходимо отметить, что сумма квадратов преобразованных данных может быть получена прибавлением $29 \times (^3/_8)$ к сумме первоначальных наблюдений пуассоновского распределения.

Литература: [1], [51, с. 54—56; русский перевод с. 108—109], [55, с. 89—90], [105, с. 187—188].

14.7. ЛОГАРИФМИЧЕСКОЕ И ОБРАТНОЕ ПРЕОБРАЗОВАНИЕ

В параграфе 14.6 мы видели, как применяется преобразование квадратного корня для получения случайной переменной, у которой среднее и дисперсия не связаны между собой, в то время как в исходной случайной величине дисперсия была пропорциональна средней. Здесь будут описаны два других часто применяемых преобразования.

Логарифмическое преобразование

$$Y = \ln X \qquad (14.7.1)$$

или

$$Y = \ln (X + 1) \qquad (14.7.2)$$

используется для получения случайных величин с несвязанными средней и дисперсией, в то время как стандартное отклонение случайной переменной X пропорционально среднему. Если коэффициент пропорциональности равен k, то дисперсия Y приближенно равна k^2. Формула (14.7.2) предпочтительнее в том случае, когда

некоторые значения X могут быть равными нулю. В качестве основания логарифма можно взять 10, тогда дисперсия будет равна $0{,}1886k^2$.

Логарифмическое преобразование применяется также в дисперсионном анализе и в регрессионных моделях, когда влияние факторов скорее мультипликативно, чем аддитивно (логарифмы аддитивны).

Обратное преобразование

$$Y = 1/X \tag{14.7.3}$$

или

$$Y = 1/(X+1) \tag{14.7.4}$$

применяется для получения случайной величины с независимыми дисперсией и средним, в то время как стандартное отклонение исходной случайной величины пропорционально квадрату среднего. Если коэффициент пропорциональности равен k, то дисперсия Y приближенно равна k^2. Формула (14.7.4) предпочтительнее в том случае, когда преобразуемые величины X могут быть равны нулю.

Так же как преобразования арксинуса и квадратного корня, логарифмическое и обратное преобразования приводят к распределениям, более близким к нормальному, чем исходное.

Все ЭВМ и большинство программируемых и непрограммируемых калькуляторов выполняют операции извлечения квадратного корня и логарифма. Таким образом, рассмотренные преобразования могут быть произведены непосредственно.

Литература: [51, с. 54—56; русский перевод с. 108—109], [97], [105, с. 182—189].

14.8. ЦЕНЗУРИРОВАННЫЕ И УСЕЧЕННЫЕ СОВОКУПНОСТИ

Пусть случайная выборка объема N извлечена из некоторой совокупности. Число наблюдений со значениями, которые меньше данного числа A и больше числа B $(B > A)$, подсчитывается, а сами наблюдения не регистрируются. Такую выборку назовем *цензурированной* снизу числом A и сверху числом B.

Цензурированные выборки встречаются довольно часто. Например, при измерении толщины сала у австралийских свиней некоторые наблюдения оказались слишком малы, а некоторые — слишком велики; для таких наблюдений может быть подсчитано общее их число. Обозначим число наблюдений, которые меньше A, через N_1, число зарегистрированных наблюдений из данного набора — через N_2, а число наблюдений со значениями больше B — через N_3. Таким образом, $N = N_1 + N_2 + N_3$.

Другой метод цензурирования состоит в отбрасывании N_1 наименьших и N_3 наибольших измерений. При этом могут быть использованы результаты из параграфа 14.9, если в качестве A взять наименьшее, а в качестве B — наибольшее из зафиксированных наблюдений. Значения N_1 и N_3 должны быть определены заранее.

В некоторых случаях N, N_1 и N_3 неизвестны. Тогда данные представляют собой множество из N_2 наблюдений, полученных из усеченного *распределения,* ограниченного снизу числом A, а сверху — числом B.

14.9. ОЦЕНИВАНИЕ ПАРАМЕТРОВ ЦЕНЗУРИРОВАННОЙ НОРМАЛЬНОЙ ВЫБОРКИ

Случайная выборка объема N извлечена из нормальной совокупности со средним μ и дисперсией σ^2. Оба параметра неизвестны. Выборка цензурирована снизу числом A и сверху числом B. Из N наблюдений значения N_1 меньше числа A, N_2 наблюдений попадают в выбранный интервал регистрации, N_3 наблюдений больше B (следовательно, $N = N_1 + N_2 + N_3$). $N2$ зарегистрированных наблюдений обозначим через $x_1, \ldots, x_{N_2}$. Вычислим два выборочных момента зарегистрированных наблюдений:

$$\overline{X} = (x_1 + \ldots + x_{N_2})/N_2, \tag{14.9.1}$$

$$S^2 = (x_1^2 + \ldots + x_{N_2}^2)/N_2 - \overline{X}^2. \tag{14.9.2}$$

Для нахождения оценок максимального правдоподобия[2] μ и σ^2 можно воспользоваться следующими рекуррентными соотношениями[3] [14]:

$$\alpha_n = (A - \mu_n)/\sigma_n, \tag{14.9.3}$$

$$\beta_n = (B - \mu_n)/\sigma_n, \tag{14.9.4}$$

$$\mu_{n+1} = \overline{X} - \sigma_n (N_1/N_2)\, \varphi(\alpha_n)/\Phi(\alpha_n) + \sigma_n (N_3/N_2) \varphi(\beta_n)/(1 - \Phi(\beta_n)), \tag{14.9.5}$$

$$\sigma_{n+1}^2 = S^2 + (\overline{X} - \mu_{n+1})^2 - \sigma_n^2 (N_1/N_2)\alpha_n \varphi(\alpha_n)/\Phi(\alpha_n) + \sigma_n^2 (N_3/N_2)\beta_n \varphi(\beta_n)/(1 - \Phi(\beta_n)). \tag{14.9.6}$$

Индекс n здесь обозначает n-е приближение оценки максимального правдоподобия, а φ и Φ — соответственно ординату и кумулятивную площадь под кривой стандартного нормального распределения. В качестве начального приближения μ_0 и σ_0^2 можно взять выборочные моменты X и S^2. Учитывая эффект усечения, лучше, возможно, в качестве σ_0^2 взять значение, несколько большее S^2, для получения μ_0 точно так же необходимо слегка подправить значение X.

Предложенные формулы несколько упрощаются, если имеется лишь одна точка усечения. Если значение этой точки больше B, то N_1 равно нулю, член, содержащий α_n, исчезает из уравнений (14.9.5) и (14.9.6) и уравнение (14.9.3) становится нам не нужным.

[2] См. параграф 13.2.

[3] Рекуррентные уравнения (14.9.5) и (14.9.6) получаются простой перестановкой уравнений максимального правдоподобия (см. параграф 3.5).

Если имеется единственная точка цензурирования A, то N_3 будет равно нулю и член, содержащий β_n, в уравнениях (14.9.5) и (14.9.6) исчезнет, соответственно уравнение (14.9.4) нам не потребуется.

Другие методы работы с цензурированными нормальными совокупностями рассматриваются в [47, с. 77—87].

Пример **14.9.1.** Из нормальной совокупности произведена случайная выборка в 30 наблюдений, цензурированных снизу числом 50 и сверху числом 150. Были получены следующие результаты:

$$N_1 = 4, \quad \overline{X} = 93{,}4,$$
$$N_2 = 25, \quad S^2 = 543{,}68,$$
$$N_3 = 1, \quad S = 23{,}316.$$

Найдем оценки максимального правдоподобия для среднего μ и дисперсии σ^2.

Ясно, что $A = 50$, а $B = 150$. В качестве начальных оценок возьмем $\mu_0 = 93$ и $\sigma_0 = 25$. Подставив эти значения в уравнения (14.9.3)—(14.9.6), получим:

$$\alpha_0 = (50 - 93)/25 = -1{,}72,$$
$$\beta_0 = (150 - 93)/25 = 2{,}28,$$
$$\mu_1 = 93{,}4 - 25\ (4/25)\ (0{,}090\ 89)/(0{,}042\ 72) + 25\ (1/25)\ \times$$
$$\times\ (0{,}029\ 65)/(0{,}011\ 30) = 87{,}51,$$
$$\sigma_1^2 = 543{,}68 + (93{,}4 - 87{,}51)^2 - (25)^2\ (4/25)\ (-1{,}72)\ \times$$
$$\times\ (0{,}090\ 89)/(0{,}042\ 72) + (25)^2\ (1/25)\ (2{,}28)\ \times$$
$$\times\ (0{,}029\ 65)/(0{,}011\ 30) = 1093{,}88.$$

Повторим вычисления, вместо μ_0 подставляя μ_1, а вместо σ_0^2 — σ_1^2. Следующие итерации приведут нас к значениям:

$$\mu_2 = 87{,}77, \quad 02 = 32{,}57,$$
$$\mu_3 = 87{,}75, \quad 0з = 32{,}97,$$
$$\mu_4 = 87{,}74, \quad \sigma_4 = 33{,}01,$$
$$\mu_5 = 87{,}75, \quad \sigma_5 = 32{,}99.$$

Заключаем, что $\hat{\mu} = 87{,}75$ и $0 = 33{,}00$.

Литература: [14], [15], [48, с. 77—78].

14.10. ОЦЕНИВАНИЕ ПАРАМЕТРОВ УСЕЧЕННОГО НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Случайная выборка объема #2 извлечена из нормального распределения, усеченного снизу A и сверху B. Значения Л и В известны, однако параметры нормального распределения μ и σ^2 неизвестны и их необходимо оценить.

Начнем с вычисления выборочных моментов X и S^2, получаемых по формулам (14.9.1) и (14.9.2). Для нахождения оценок

максимального правдоподобия[4] μ и σ^2 можно воспользоваться следующими рекуррентными соотношениями[5] [14]:

$$\alpha_n = (A - \mu_n)/\sigma_n, \tag{14.10.1}$$
$$\beta_n = (B - \mu_n)/\sigma_n, \tag{14.10.2}$$
$$\mu_{n+1} = \overline{X} + \sigma_n\ (\phi\ (\beta_n) - \varphi\ (\alpha_n))/(\Phi\ (\beta_n) - \Phi\ (\alpha_n)), \tag{14.10.3}$$
$$\sigma_{n+1}^2 = S^2 + (X - \mu_{n+1})^2 + \sigma_n^2\ (\beta_n \varphi\ (\beta_n) -$$
$$- \text{ад»}\ (\alpha_n))/(\Phi\ (\beta_n) - \Phi\ (\alpha_n)). \tag{14.10.4}$$

Индекс n обозначает n-е приближение оценки максимального правдоподобия, а ϕ и Φ — соответственно ординату и кумулятивную площадь под кривой стандартного нормального распределения. Выборочные моменты X и S^2 снова могут быть использованы в качестве начальных приближений μ_0 и σ_0^2. Вероятно, значение σ_0^2, которое больше S^2, в силу эффекта усечения будет лучшей аппроксимацией. Необходимо аналогично несколько подправлять и $\overline{X}$.

Формулы (14.10.3) и (14.10.4) упростятся, если имеется лишь одна точка усечения. Если значение этой точки равно B, то члены $\varphi(\alpha_n)$, $\alpha_n\varphi(\alpha_n)$ и $\Phi(\alpha_n)$ исчезнут и уравнение (14.10.1) нам не потребуется. Если имеется единственное цензурированное наблюдение в точке A, то члены $\phi\ (\beta_n)$ и $\beta_n\varphi\ (\beta_n)$ исчезают, $\Phi\ (\beta_n)$ равно 1 и уравнение (14.10.2) нам не потребуется.

Иногда точки цензурирования A и B неизвестны. В этом случае, если объем выборки достаточно велик, в качестве A можно взять значение, несколько меньшее наименьшего из наблюдений, а в качестве B — значение, несколько большее наибольшего.

Другие методы в условиях усеченных нормальных совокупностей описаны в [48, с. 77—78].

Пример 14.10.1. Из нормального распределения извлечена случайная выборка в 25 наблюдений, усеченная снизу числом 50 и сверху числом 150. Получены следующие результаты: $X = 93{,}4$; $S^2 = 543{,}68$. Найдем оценки максимального правдоподобия параметров μ И σ^2.

Снова имеем $A = 50$ и $B = 150$. В качестве начальной оценки положим $\mu_0 = 93$ и $\sigma_0 = 25$. Подставив эти значения в уравнения (14.10.1)—(14.10.4), получим:

$$\alpha_0 = (50 - 93)/25 = -1{,}72,$$
$$\beta_0 = (150 - 93)/25 = 2{,}28,$$
$$\mu_1 = 93{,}4 + 25\ (0{,}029\ 65 - 0{,}090\ 89)/(0{,}988\ 70 - 0{,}042\ 72) = 91{,}78,$$
$$\sigma_1^2 = 543{,}68 + (93{,}4 - 91{,}78)^2 + 25^2\ \{(2{,}28)\ (0{,}029\ 65) -$$
$$- (-1{,}72)\ (0{,}090\ 89)\} \div (0{,}988\ 70 - 0{,}042\ 72) = 694{,}25.$$

[4] См. параграф 13.2.

[5] Рекуррентные уравнения (14.10.3) и (14.10.4) получаются простой перестановкой уравнений максимального правдоподобия (см. параграф 3.5).

Повторим вычисления, подставив μ_1 вместо $_0$ и σ_1^2 вместо a. Итерации приведут нас к следующим значениям:

$$\mu_2 = 91{,}10, \quad \sigma_2 = 27{,}21,$$
$$\mu_3 = 90{,}76, \quad \sigma_3 = 27{,}77,$$
$$\mu_4 = 90{,}51, \quad \sigma_4 = 28{,}16,$$
$$\mu_5 = 90{,}33, \quad \sigma_5 = 28{,}47,$$
$$\mu_6 = 90{,}19, \quad \sigma_6 = 28{,}69.$$

Сходимость достаточно медленная. Приближения $\{\mu_n\}$ равномерно убывают, а $\{\sigma_n\}$ возрастают. Для надежности выполним еще несколько итераций:

$$\mu_6 = 89{,}00, \quad 06 = 30{,}00,$$
$$\mu_7 = 89{,}29, \quad \sigma_7 = 29{,}88,$$
$$\mu_8 = 89{,}44, \quad 08 = 29{,}77.$$

Теперь $\{\mu_n\}$ возрастают, а $\{\sigma_n\}$ убывают. Проведем еще три итерации:

$$\mu_8 = 89{,}70, \quad 08 = 29{,}33,$$
$$\mu_9 = 89{,}74, \quad \sigma_9 = 29{,}34,$$
$$\mu_{10} = 89{,}74, \quad \sigma_{10} = 29{,}35.$$

Заключаем, что $\hat{\mu} = 89{,}74$ и $\hat{a} = 29{,}35$. Следует отметить большое количество потребовавшихся арифметических вычислений. Окончательный результат можно сравнить с примером 14.9.1, в котором была использована дополнительная информация.

Литература: [14], [15], [22], [48, с. 77—87].

14.11. ОЦЕНИВАНИЕ ПАРАМЕТРОВ ЦЕНЗУРИРОВАННОИ ПУАССОНОВСКОЙ ВЫБОРКИ

Методы оценивания параметров цензурированных пуассоновских распределений описаны в [47, с. 104—109]. Однако если параметр λ не слишком мал, то можно воспользоваться модификацией метода из параграфа 14.9. Достаточно точные результаты могут быть получены, если K приблизительно больше 8 или 9.

Допустим, что выборка объема N извлечена из совокупности с пуассоновским распределением со средним λ, которое неизвестно. Наблюдения, значения которых меньше a и больше $Ь$, подсчитываются, но не регистрируются, как a, так и b — целые числа, $b > a$. Из N наблюдений N_1 меньше a, N_2 попадают в выбранный интервал регистрации, N_3 больше b (следовательно, $N = N_1 + N_2 + N_3$). Обозначим N_2 зарегистрированных наблюдений через $x_1, \ldots, x_{N_2}$. По формуле (14.9.1) найдем выборочное среднее X.

Положим $A = a - {}^1/_2$ и $B = b + {}^1/_2$. Приближенную оценку максимального правдоподобия [6] для λ можно найти из следующих уравнений:

$$\alpha_n = (A - \lambda_n)/\sqrt{\lambda_n}, \tag{14.11.1}$$
$$\beta_n = B - \lambda_n)/\sqrt{\lambda_n}, \tag{14.11.2}$$
$$\lambda_{n+1} = \overline{X} - \lambda_n^{\frac{1}{2}} (N_1/N_2)\, \varphi(\alpha_n)/\Phi(\alpha_n) + \lambda_n^{\frac{1}{2}} (N_3/N_2)\, \varphi(\beta_n)/(1 - \Phi(\beta_n)). \tag{14.11.3}$$

Индекс n обозначает n-е приближение к оценке максимального правдоподобия, а φ и Φ — соответственно ординату и кумулятивную площадь под кривой стандартного нормального распределения. Уравнение (14.11.3) получено на основе аппроксимации пуассоновского распределения (см. параграф 10.8) нормальным. Оно вытекает из (14.9.5) при подстановке λ_n вместо μ_n и σ^2. В качестве начального приближения можно положить $\lambda_0 = X$. Если N_3 намного больше N_1, то в качестве λ_0 возьмем значение, несколько большее X. Тогда скорость сходимости увеличится. Обратно: если N_1 намного больше N_3, то в качестве λ_0 выберем значение, несколько меньшее X.

Приведенные формулы упростятся, если имеется всего одна точка цензурирования. Если эта точка больше b, то N_1 равно нулю, член, содержащий α_n, исчезает из уравнения (14.11.3) и уравнение (14.11.1) нам не потребуется. Если точка цензурирования меньше a, то N_3 равно нулю, член, содержащий β_n, исчезает из уравнения (14.11.2).

Пример **14.11.1.** Из распределения Пуассона с неизвестным средним λ извлечена случайная выборка объема 29. Пять наблюдений больше 12, а среднее остальных наблюдений равно 8,125. Найдем оценку максимального правдоподобия параметра Я.

В этом примере имеется только одна верхняя точка цензурирования и мы используем уравнение (14.11.3) при $N_1 = 0$. Заметим, что $B = 12{,}5$. В качестве начального приближения выберем $Ko = \overline{X} = 8{,}125$. Найдем

$$\text{Po} = (12{,}5 - 8{,}125)/(8{,}125)^{1/2} = 1{,}5349.$$

Приближение λ_1 получаем из уравнения (14.11.3):

$$\lambda_1 = 8{,}125 + (8{,}125)^{1/2}(5/24)\,(0{,}122\,84)/(0{,}062\,41) = 9{,}294.$$

Теперь можно вычислить β_1 и λ_2. В результате итераций получаем:

$$\lambda_2 = 9{,}120; \quad \lambda_3 = 9{,}145; \quad \lambda_4 = 9{,}141; \quad \lambda_5 = 9{,}142; \quad \lambda_6 = 9{,}142.$$

Приближенная оценка максимального правдоподобия есть $\hat{\lambda} = 9{,}142$.

Литература: [47, с. 104—109].

[6] См. параграф 13.2.

14.12. ОЦЕНИВАНИЕ ПАРАМЕТРОВ УСЕЧЕННОГО ПУАССОНОВСКОГО РАСПРЕДЕЛЕНИЯ

Методы оценивания параметров усеченного пуассоновского распределения описаны в [47, с. 104—109]. Если параметр K не слишком мал, то можно воспользоваться модификацией метода из параграфа 14.10. Достаточно точные результаты можно получить для значений Я, которые приблизительно больше 8 или 9.

Допустим, случайная выборка объема NZ извлечена из усеченного пуассоновского распределения с интервалом регистрации $[a, b]$ (a и b — целые известные константы, $b > a$). Параметр пуассоновского распределения Я известен и его необходимо оценить. Значения выборки есть $x_1, x_2, \ldots, x_{N_0}$.

Определим $A = a - 1/2$ и $B = b + 1/2$ и найдем выборочное среднее X по формуле (14.9.1). Приближенную оценку максимального правдоподобия Я можно получить с помощью следующих рекуррентных соотношений:

$$\alpha_n = (\text{Л} - \lambda_n)/\sqrt{\lambda_n}, \qquad (14.12.1)$$

$$\beta_n = (B - \lambda_n)/\sqrt{\lambda_n}, \qquad (14.12.2)$$

$$\lambda_{n+1} = X + \lambda_n^{\frac{1}{2}} (\text{ср}\ (\beta_n) - \varphi(\alpha_n))/(\Phi(\beta_n) - \Phi(\alpha_n)). \qquad (14.12.3)$$

Индекс n обозначает n-е приближение к оценке максимального правдоподобия, а φ и Φ — соответственно ординату и кумулятивную площадь под кривой стандартного нормального распределения. Уравнение (14.12.3) получено на основе аппроксимации пуассоновского распределения нормальным (см. параграф 10.8). Оно вытекает из уравнения (14.10.3) при подстановке λ_n вместо μ_n и σ^2. В качестве начального приближения можно взять $\lambda_0 = X$.

Приведенные формулы упростятся, если имеется всего одна точка усечения. Если эта точка верхняя и она равна b, то $\varphi(\alpha_n)$ и $\Phi(\alpha_n)$ принимают значение нулей и уравнение (14.12.2) нам не потребуется.

Пример 14.12.1. Из пуассоновского распределения извлечена выборка объема 24, усеченная неотрицательными числами 0 и 12. Выборочное среднее $\bar{X} = 8{,}125$. Получим оценку максимального правдоподобия параметра Я.

В данном случае имеется одна точка усечения (верхняя), поэтому в уравнении (14.12.3) значения $\varphi(\alpha_n)$ и $\Phi(\alpha_n)$ равны нулю.

B по-прежнему равно 12,5. В качестве начального приближения положим $\lambda_0 = X = 8{,}125$. Далее вычислим:

$$\beta_0 = (12{,}5 - 8{,}125)/(8{,}125)^{1/2} = 1{,}5349,$$

$$\lambda_1 = 8{,}125 + (8{,}125)^{1/2}(0{,}122\,84)/(0{,}937\,59) = 8{,}498.$$

Найдем β_1 и повторим процесс. Итерации приведут нас к следующим значениям:

$$\lambda_2 = 8{,}620; \quad \lambda_3 = 8{,}665; \quad \lambda_4 = 8{,}681; \quad \lambda_5 = 8{,}637; \quad \lambda_6 = 8{,}690;$$
$$\lambda_7 = 8{,}631; \quad \lambda_8 = 8{,}691.$$

Таким образом, приближенная оценка максимального правдоподобия есть $\hat{\text{Я}} = 8{,}691$. Этот результат предлагаем читателю сравнить с результатом, полученным в примере 14.11.1, где используется дополнительная информация.

Возможно, стоит отметить, что выборка извлечена из пуассоновской совокупности со средним, равным 9. Данные представлены в табл. 14.6.1.

Литература: [47, с. 104—109].

14.13. УПРАЖНЕНИЯ

1. Опишите процедуру генерирования случайных наблюдений из t-распределения с 3 степенями свободы.

2. Треугольное распределение на единичном интервале (О, 1) задается функцией плотности вероятностей $f(x) = 2x$ или кумулятивной функцией распределения $F(x) = x^2$. Опишите процедуру получения случайных наблюдений с этим распределением.

3. В биномиальном эксперименте с 18 испытаниями наблюдалось 14 случаев «успеха». С помощью преобразования арксинуса (14.5.1) найдите 95 %-ный доверительный интервал для p (вероятность успеха в одном испытании). Сравните свой ответ с результатами примера 13.5.1.

4. В табл. 14.13.1 представлены данные о толщине сала 387 австралийских свиней. На основе этих данных постройте нормальную кривую.

Таблица 14.13.1. **Толщина сала 387 австралийских свиней**

Толщина сала, мм	Число	Толщина сала, мм	Число
$\leqslant 6$	46	15	13
7	25	16	14
8	32	17	7
9	50	18	5
10	37	19	8
11	50	20	0
12	44	21	2
13	33	$\geqslant 22$	0
14	21		
		Итого	387

Источник. Сообщено в частной беседе Х. К. Киртоном, штат Новый Южный Уэльс, Сидней, Австралия, Министерство сельского хозяйства.

Часть III. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

15. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ И МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

В параграфах 15.1 и 15.2 этой главы рассматривается техника вычислений, связанная с применением метода наименьших квадратов при подборе прямой к множеству точек на плоскости xy. В параграфах 15.4 и 15.5 описываются стандартные статистические процедуры при парной линейной регрессии и соответствующие критерии значимости. Методы оценивания свободного члена β_0, коэффициента наклона β_1 и дисперсии ошибки σ^2 обсуждаются в параграфах 15.6 и 15.7. Критерий значимости параметров регрессии разбирается в параграфе 15.8. В параграфе 15.10 будут описаны методы определения среднего значения y по данному значению x. Качество подгонки, неравные дисперсии и взвешенный метод наименьших квадратов будут рассмотрены в параграфах 15.11 и 15.12. Матричные методы являются излишними при описании парной линейной зависимости, несмотря на это матричный подход будет освещен в параграфах 15.3 и 15.9 в целях подготовки читателя к гл. 16 и 17.

15.1. ВВЕДЕНИЕ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Предположим, через четыре точки на плоскости, изображенные на рис. 15.1.1, $(x_1, y_1) = (1, 10)$, $(x_2, y_2) = (3, 20)$, $(x_3, y_3) = (4, 18)$, $(x_4, y_4) = (5, 20)$, требуется провести прямую так, чтобы она прошла как можно ближе к ним. Эти точки на рисунке обозначены крестиками. Первый способ заключается в том, чтобы, расположив линейку как можно ближе к этим точкам, провести по ней прямую линию. Однако что мы понимаем под словом «близко» и как узнать, что прямая прошла «как можно ближе»? Даже в простой ситуации, изображенной на рис. 15.1.1, лучшее положение прямой является далеко не очевидным.

В *методе наименьших квадратов* используется критерий, который приводит к единственному решению. Метод может быть объяснен следующим образом. Рассмотрим прямую линию, проходящую близко к заданным точкам так, как это изображено на рис. 15.1.1. Если мы примем изображенную ка этом рисунке прямую за искомую, то подобранные значения y в точках $x = 1$, $x = 3$, $x = 4$ и $x - 5$ обозначим маленькими кружками. Подгонка (аппроксимация) считается хорошей, если все расстояния между наблюдаемыми значениями (обозначенные крестиками) и соответствующими расчетными предсказанными (обозначенные кружками) достаточно малы. Прямая, найденная по методу наименьших квадратов, минимизирует сумму квадратов этих расстояний.

Метод наименьших квадратов как вычислительная процедура был описан Лагранжем в 1806 г. в его труде Nouvelles méthodes pour la determination des orbites des comètes. Им также было предложено название этого метода. Первым, кто связал метод наименьших квадратов с теорией вероятностей, был Гаусс (1809 г.)

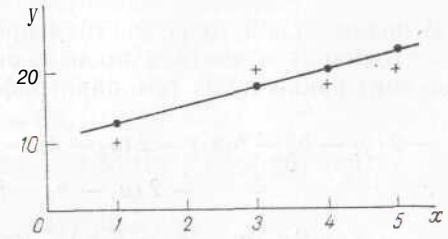


Рис. 15.1.1. Прямая линия, построенная по четырем точкам. Точки наблюдений отмечены крестиками, расчетные точки — кружками

[29]. Гаусс заметил, что этот метод применялся им еще с 1795 г. Применял его и Лаплас, но совершенно в других целях (1811 г.) (см. [80], [100, с. 209—228]).

Термин *регрессия* был введен Фрэнсисом Гальтоном в 1886 г. Гальтон обнаружил, что в среднем сыновья высоких отцов имеют не такой большой рост, а сыновья отцов с небольшим ростом выше своих отцов. Это было интерпретировано им как «регрессия к посредственности». Ошибки в рассуждениях Гальтона были разъяснены, например, Браунли [7, с. 407, русский перевод с. 375—376] (см. также [45, с. 201]).

Литература: [3, с. 185—187], [8, с. 31—32, 272—275], [19, с. 1—6; русский перевод с. 9—15], [30]—[33], [45, с. 195—196], [50, с. 377—381], [66, с. 116—120].

15.2. ПОДБОР ПРЯМОЙ ЛИНИИ[1] ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ

Обозначим расчетное (выравненное) значение y в точке x_i через Y_i и положим

$$Y_i = b_0 + b_1 x_i, \tag{15.2.1}$$

где b_0 и b_1 — константы, подлежащие определению. При применении метода наименьших квадратов мы хотим минимизировать

$$\sum_{i=1}^{4} (y_i - Y_i)^2. \tag{15.2.2}$$

[1] Если требуется, чтобы прямая линия прошла через центр или другую фиксированную точку, необходимо применить метод из параграфа 16.11.

Это значит необходимо минимизировать[2]

$$\sum_{i=1}^{4}(y_i - b_0 - b_1 x_i)^2, \qquad (15.2.3)$$

подбирая значения b_0 и b_1 (значения x и y нам известны). Критерий (15.2.3) может быть переписан следующим образом:

$$(y_1 - b_0 - b_1x_1)^2 + (y_2 - b_0 - b_1x_2)^2 + (y_3 - b_0 - b_1x_3)^2 + (y_4 - b_0 - b_1x_4)^2.$$

В оптимальной точке частная производная по b_0 (если считать b_1 постоянной) и частная производная по b_1 (если считать b_0 постоянной) равны нулю (см. параграф 1.6). Тогда

$$-2(y_1 - b_0 - b_1x_1) - 2(y_2 - b_0 - b_1x_2) - 2(y_3 - b_0 - b_1x_3) - 2(y_4 - b_0 - b_1x_4) = 0,$$

$$-2x_1(y_1 - b_0 - b_1x_1) - 2x_2(y_2 - b_0 - b_1x_2) - 2x_3(y_3 - b_0 - b_1x_3) - 2x_4(y_4 - b_0 - b_1x_4) = 0.$$

Эти уравнения могут быть преобразованы:

$$\left.\begin{aligned} 4b_0 + \left(\sum_{i=1}^{4} x_i\right) b_1 &= \sum_{i=1}^{4} y_i, \\ \left(\sum_{i=1}^{4} x_i\right) b_0 + \left(\sum_{i=1}^{4} x_i^2\right) b_1 &= \sum_{i=1}^{4} x_i y_i. \end{aligned}\right\} \qquad (15.2.4)$$

Решим эти два линейных уравнения *(нормальные уравнения)* для получения b_0 и b_1. В общем случае для n точек нормальные уравнения принимают следующий вид:

$$\left.\begin{aligned} nb_0 + \left(\sum_{i=1}^{n} x_i\right) b_1 &= \sum_{i=1}^{n} y_i, \\ \left(\sum_{i=1}^{n} x_i\right) b_0 + \left(\sum_{i=1}^{n} x_i^2\right) b_1 &= \sum_{i=1}^{n} x_i y_i. \end{aligned}\right\} \qquad (15.2.5)$$

Существует простое решение этой системы:

$$\left.\begin{aligned} b_1 &= \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \Big/ \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right), \\ b_0 &= \bar{y} - b_1\bar{x}. \end{aligned}\right\} \qquad (15.2.6)$$

[2] В общем случае имеем га точек и суммирование в (15.2.3) ведется от 1 до n.

Необходимо отметить, что прямая метода наименьших квадратов проходит через среднюю точку (x, y); тогда уравнение прямой можно записать так:

$$Y = \bar{y} + r(s_y/s_x)(x - \bar{x}), \qquad (15.2.7)$$

где s_x и s_y обозначают стандартные отклонения x и y, а r — коэффициент корреляции. Формулу (15.2.7) предлагаем читателю сравнить с формулой (9.6.2).

Пример **15.2.1.** Для четырех точек (см. параграф 15.1) находим:

$$\sum x_i = 1 + 3 + 4 + 5 = 13,$$

$$\sum x_i^2 = 1 + 9 + 16 + 25 = 51,$$

$$\sum y_i = 10 + 20 + 18 + 20 = 68,$$

$$\sum x_i y_i = (1 \times 10) + (3 \times 20) + (4 \times 18) + (5 \times 20) = 242.$$

Нормальные уравнения имеют следующий вид:

$$4b_0 + 13b_1 = 68,$$

$$13b_0 + 51b_1 = 242.$$

Решая эти уравнения, найдем $b_0 = 9{,}2$ и $b_1 = 2{,}4$. Таким образом, подогнанная прямая имеет вид: $Y = 9{,}2 + 2{,}4x$. Расчетные значения Y равны:

$$Y_1 = 9{,}2 + 2{,}4 \times 1 = 11{,}6,$$

$$Y_2 = 9{,}2 + 2{,}4 \times 3 = 16{,}4,$$

$$Y_3 = 9{,}2 + 2{,}4 \times 4 = 18{,}8,$$

$$Y_4 = 9{,}2 + 2{,}4 \times 5 = 21{,}2.$$

Как правило, коэффициенты b_0 и b_1 вычисляются непосредственно по формулам (15.2.6). Тогда

$$b_1 = \{242 - 4 \times (13/4)\,(68/4)\}/\{51 - 4 \times (13/4)^2\} = 2{,}4,$$

$$b_0 = (68/4) - 2{,}4 \times (13/4) = 9{,}2.$$

Литература: [3, с. 185—187], [19, с. 7—12; русский перевод с. 15—22], [26, с. 296—304], [50, с. 382—398; русский перевод с. 447—458], [65, с. 109—128], [66, с. 121—130], [86, с. 254—256, 260—264], [92, с. 536—543], [93, с. 217—240], [102, с. 240—244].

15.3. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ. МАТРИЧНОЕ ОБОЗНАЧЕНИЕ

Для примера 15.2.1 с парной линейной регрессией матричное обозначение вряд ли является необходимым, однако с его помощью можно обобщить полученные результаты для расчета криволиней-

ной и множественной регрессии. Определим в нашем примере с четырьмя точками матрицу X, векторы y, Y и b следующим образом:

$$\mathbf{X}=\begin{pmatrix}1 & x_1\\ 1 & x_2\\ 1 & x_3\\ 1 & x_4\end{pmatrix},\quad \mathbf{y}=\begin{pmatrix}y_1\\ y_2\\ y_3\\ y_4\end{pmatrix},\quad \mathbf{Y}=\begin{pmatrix}Y_1\\ Y_2\\ Y_3\\ Y_4\end{pmatrix},\quad \mathbf{b}=\begin{pmatrix}b_0\\ b_1\end{pmatrix}.$$

Заметим теперь, что

$$\mathbf{Xb}=\begin{pmatrix}b_0+b_1x_1\\ b_0+b_1x_2\\ b_0+b_1x_3\\ b_0+b_1x_4\end{pmatrix}\quad \text{и}\quad \mathbf{y}-\mathbf{Xb}=\begin{pmatrix}y_1-b_0-b_1x_1\\ y_2-b_0-b_1x_2\\ y_3-b_0-b_1x_3\\ y_4-b_0-b_1x_4\end{pmatrix},$$

поэтому критерий, по которому минимизируется сумма (15.2.3), может быть переписан как

$$(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}). \qquad (15.3.1)$$

Расчетные значения, задаваемые формулой (15.2.1), могут быть записаны так:

$$\mathbf{Y}=\mathbf{Xb}. \qquad (15.3.2)$$

Исследуем матрицу $\mathbf{S}=\mathbf{X}'\mathbf{X}$ и вектор $\mathbf{X}'\mathbf{y}$. Как видим, S имеет порядок 2X2, а размерность вектор-столбца $\mathbf{X}'\mathbf{y}$ равна 2. Таким образом,

$$\mathbf{S}=\begin{pmatrix}4 & \sum_{i=1}^{4}x_i\\ \sum_{i=1}^{4}x_i & \sum_{i=1}^{4}x_i^2\end{pmatrix}\quad \text{и}\quad \mathbf{X}'\mathbf{y}=\begin{pmatrix}\sum_{i=1}^{4}y_i\\ \sum_{i=1}^{4}x_iy_i\end{pmatrix}.$$

Поэтому нормальные уравнения (15.2.4) в матричном виде могут быть записаны следующим образом:

$$\mathbf{Sb}=\mathrm{X'y}, \qquad (15.3.3)$$

а их решение — как

$$\mathbf{b}=\mathbf{S}^{-1}\mathbf{X}'\mathbf{y}. \qquad (15.3.4)$$

Приведенные выше матрицы и векторы определяются точно так же и для общего случая с n точками. Далее мы увидим, что к такой же системе линейных уравнений (15.3.3) приводят криволинейные и множественные регрессии; размерность матрицы X тогда будет больше, чем $n\times 2$.

Литература: [19, с. 44—52; русский перевод с. 53—62].

15.4. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ. СТАТИСТИЧЕСКАЯ МОДЕЛЬ

В параграфах 15.1—15.3 метод наименьших квадратов был рассмотрен с вычислительной точки зрения. Этот метод обычно применяется для экспериментальных данных. Экспериментальные наблюдения, однако, подвержены статистическим колебаниям: при повторении эксперимента мы получим другие наблюдения, поэтому результаты применения метода наименьших квадратов будут не одинаковы.

Для статистического анализа необходимо построение математической модели. Предположим

$$y_i=\beta_0+\beta_1x_i+e_i, \qquad (15.4.1)$$

где Ро и β_1 — неизвестные константы, а ошибки $\{e_i\}$ — независимые случайные величины, имеющие нормальное распределение с нулевым математическим ожиданием и одинаковой неизвестной дисперсией σ^2; x рассматривается как *независимая переменная* и считается, что эта переменная не содержит ошибок, а y — *зависимая переменная.* Значения *bo* и b_1, которые мы вычисляем, являются оценками β_0 и PI и основываются на имеющихся результатах эксперимента. Если β_1 равно нулю, то линейная связь между x и y отсутствует.

Литература:[7, с. 334—337; русский перевод с. 309—312], [19, с. 17; русский перевод с. 18], [86, с. 257—259].

15.5. КРИТЕРИЙ ЗНАЧИМОСТИ ЛИНИИ РЕГРЕССИИ

Даже в том случае, когда x и y независимы, наблюдаемые значения можно нанести на плоскость в виде точек и подобрать к ним прямую по методу наименьших квадратов. Такая прямая будет иметь, как правило, нулевой коэффициент наклона. Ненулевой наклон является тогда следствием случайности и не отражает линейной зависимости между x и y. Как проверить, является ли регрессия значимой?

Начнем с исследования общей суммы квадратов отклонений значений $\{y_i\}$ от среднего y:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2. \qquad (15.5.1)$$

Можно показать, что для метода наименьших квадратов имеет место следующее разложение:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(Y_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-Y_i)^2. \qquad (15.5.2)$$

Таким образом, сумма квадратов (15.5.1) может быть разбита на две положительные компоненты:

а) сумму квадратов значений регрессии относительного среднего;

б) сумму квадратов отклонений относительно линии регрессии (остаточная сумма квадратов).

Сумма квадратов регрессии есть сумма квадратов разности между значениями, найденными на основе регрессии, и средним. Сумма квадратов относительно линии регрессии есть сумма квадратов расстояний между наблюдаемыми точками и точками, полученными на основе регрессии (15.5.2). Если подобранная прямая проходит через все имеющиеся точки, она является идеальной и сумма квадратов отклонений относительно этой прямой будет равна нулю, а вся вариация значений $\{y_i\}$ *объясняется* прямой. С другой стороны, если данные не содержат линейного тренда, то сумма квадратов регрессии относительно среднего будет мала и почти вся вариация в изменении $\{y_i\}$ может быть объяснена как вариация относительно линии регрессии.

Поэтому представляется, что регрессия будет значимой, если сумма квадратов регрессии относительно среднего *будет больше* по сравнению с суммой квадратов отклонений относительно линии регрессии.

Формально нам необходимо проверить нулевую гипотезу H_0: $\beta_1 = 0$ против альтернативы H_1: $\beta_1 \neq 0$. Если мы установим, что β_1 не равно нулю, то регрессию называем значимой. Вычисления по проверке значимости регрессии лучше всего проводить в так называемой *таблице дисперсионного анализа*. В этой таблице общая сумма квадратов разбита на сумму квадратов регрессии и остаточную сумму квадратов (сумма квадратов отклонений относительно линии регрессии), как показано в (15.5.2). Общепринятые формулы [3] вычислений приведены в табл. 15.5.1 [4].

Таблица 15.5.1. **Дисперсионный анализ парной линейной регрессии**

Источник вариации (1)	с. к. (2)	с. с. (3)	ср. к. (4) = (2)/(3)
Регрессия	$b_1^2\left(\sum_1^n x_i^2 - n\bar{x}^2\right)$	1	(определяется делением)
Остаток	(определяется вычитанием)	$n-2$	
Общая вариация	$\sum_1^n y_i^2 - n\bar{y}^2$	$n-1$	—

Суммы квадратов регрессии и отклонений в столбце (4) получают делением суммы квадратов из столбца (2) на соответствующую степень свободы в столбце (3).

[3] Можно привести и более подробные математические формулы.

[4] В таблицах «сумма квадратов», «степень свободы» и «средний квадрат» сокращенно обозначены как с. к., с. с., ср. к.

Вычислим далее

$$F_{1,n-2} = \frac{\text{средний квадрат регрессии}}{\text{средний квадрат отклонений}}. \qquad (15.5.3)$$

Для нулевой гипотезы $\beta_1 = 0$ эта статистика имеет F-распределение. Если вычисленное значение достаточно велико, то нулевая гипотеза отклоняется и мы делаем вывод, что β_1 не равно нулю, следовательно, регрессия значима.

F-критерий в парной линейной регрессии эквивалентен более часто применяемому двустороннему t-критерию:

$$t_{n-2} = (n-2)^{\frac{1}{2}} r/(1-r^2)^{\frac{1}{2}}, \qquad (15.5.4)$$

где r — коэффициент корреляции между x и y (см. параграф 8.8). Действительно, F-статистика в парной линейной регрессии есть квадрат t-статистики (см. параграф 9.4). Формула (15.5.4) совпадает с критериальной статистикой, применяемой для проверки значимости коэффициента корреляции (12.31.2).

Пример 15.5.1. Проверим значимость регрессии из примера 15.2.1. Вычислим:

$$\sum y_i^2 = 10^2 + 20^2 + 18^2 + 20^2 = 1224,$$

$$\text{с. к. регрессии} = (2{,}4)^2\{51 - 4(13/4)^2\} = 50{,}4,$$

$$\text{общая с. к.} = 1224 - 4(68/4)^2 = 68{,}0,$$

$$\text{остаточная с. к.} = 68{,}0 - 50{,}4 = 17{,}6.$$

Заполним таблицу дисперсионного анализа 15.5.2 и вычислим: $F_{1,2} = 50{,}4/8{,}8 = 5{,}73$. Это значение не значимо при 5 %-ном уровне, поэтому у нас нет доказательств в пользу того, что β_1 равно нулю.

Таблица 15.5.2. **Дисперсионный анализ парной линейной регрессии из примера 15.5.1**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	50,4	1	50,4
Остаток	17,6	2	8,8
Общая вариация	68,0	3	—

Доля общей суммы квадратов, объясняемой регрессией, называется *коэффициентом детерминации* и обозначается R^2. Это значение равно квадрату коэффициента корреляции между наблюдаемыми и вычисленными на основе регрессии значениями y *(коэффициент множественной корреляции)* [5]. Как видно из табл. 15.5.2,

$$R^2 = 50{,}4/68{,}0 = 0{,}7412.$$

[5] См. параграф 17.2. В случае парной линейной регрессии $R^2 = r^2$.

В этом примере сумма квадратов была найдена по обычной формуле (см. табл. 15.5.1). Читатель может проверить, что то же значение суммы квадратов можно получить, воспользовавшись формулой (15.5.2).

Литература: [3, с. 188—193], [19, 24—26; русский перевод с. 22—24], [105, с. 205—209].

15.6. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ

Параметры β_0, β_1 и σ^2 неизвестны, однако можно найти их несмещенные оценки. Коэффициенты регрессии b_0 и b_1 оценивают параметры β_0 и Рь а средний квадрат отклонений в таблице дисперсионного анализа оценивает σ^2. Формулы дисперсий и ковариаций b_0 и b_1 приводятся в параграфе 15.9.

15.7. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ [6] ДЛЯ Ро, β_1 И σ^2

Обозначим средний квадрат отклонений в таблице дисперсионного анализа через s^2. 95 %-ный доверительный интервал для Ро и β_1 может быть найден приравниванием t-статистик к верхней и нижней 2,5 %-ным точкам t_{n-2}-распределения:

$$t'_{n-2} = \frac{b_0 - \text{Ро}}{s\{(\sum x_i^2)/(n\sum x_i^2 - n^2\bar{x}^2)\}^{\frac{1}{2}}}, \qquad \textbf{(15.7.1)}$$

$$t_{n-2} = \frac{b_1 - \beta_1}{s\{1/(\sum x_i^2 - n\bar{x}^2)\}^{\frac{1}{2}}}. \qquad (15.7.2)$$

Для нахождения 95 %-ного доверительного интервала для σ^2 приравняем статистику χ^2 к верхней и нижней 2,5 %-ным точкам распределения χ^2_{n-2}:

$$\chi^2_{n-2} = \frac{(n-2)\,s^2}{\sigma^2}. \qquad (15.7.3)$$

Пример **15.7.1.** Найдем 95 %-ные доверительные границы для свободного члена β_0 и дисперсии ошибки σ^2 по данным табл. 15.7.1.

Таблица 15.7.1. **Данные для расчета регрессии в примере 15.7.1.**

x	y	x	y	x	y	x	y
1	17	6	28	11	44	16	60
2	13	7	26	12	47	17	58
3	22	8	28	13	45	18	61
4	20	9	34	14	54	19	64
5	20	10	45	15	55	20	70

[6] Определения доверительных интервалов см. в параграфе 13.4.

Начнем с вычисления следующих величин:

$\bar{x} = 10{,}5$, $\sum x_i^2 = 2870$, $\sum x_i y_i = 10503$, $b_0 = 9{,}384\,21$,

$\bar{y} = 40{,}6$ $\sum y_i^2 = 390\,50$, $n = 20$, $b_1 = 2{,}972\,93$.

Заполним клетки табл. 15.7.2.

Таблица 15.7.2. **Дисперсионный анализ данных из примера 15.7.1**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	5877,437	1	5877,487
Остаток	205,313	18	11,406
Общаявариация	6082,800	19	—

Для нахождения 95 %-ных доверительных границ для Ро решим уравнения

$$\frac{9{,}38421 - \beta_0}{[11{,}406 \times 2870/\{20 \times 2870 - 20^2 \times (10{,}5)^2\}]^{1/2}} = \pm 2{,}101,$$

которые приводят нас к интервалу 9,384 + 3,296.

Для нахождения 95 %-ных доверительных границ для σ^2 решим уравнения

$$\frac{18 \times 11{,}406}{\sigma^2} = 31{,}53 \text{ и } \frac{18 \times 11{,}406}{\sigma^2} = 8{,}23,$$

которые дают доверительные границы 6,51 и 24,95.

Литература: [19, с. 18—21; русский перевод с. 27—28], [54, с. 362—363], [86, с. 265—272], [105, с. 209—210].

15.8. ПРОВЕРКА ГИПОТЕЗ О ПАРАМЕТРАХ Ро, β_1 И σ^2

F- и *t*-методы проверки нулевой гипотезы $\beta_1 = 0$ были описаны в параграфе 15.5. Предположим теперь, что необходимо проверить нулевую гипотезу $\beta_1 = 6$, где δ — некоторая известная константа. Проверка может быть осуществлена несколькими способами, читатель может воспользоваться по своему усмотрению одним из них, все они математически эквивалентны.

Один способ заключается в том, чтобы приравнять в выражении (15.7.2) β_1 к *8*. Мы отклоняем нулевую гипотезу $\beta_1 = 8$, если *t*-статистика попадает в верхнюю или нижнюю 2,5 %-ную область t_{n-2}-распределения. Другой способ заключается в том, чтобы, вычитая из y_i δx_i для каждого *i*, найти z_i и по точкам $\{(x_i, z_i)\}$ выполнить проверку по /- или *F*-критерию из параграфа 15.5. *t*-метод можно также применять для проверки односторонних гипотез.

Для проверки нулевой гипотезы Ро = б, где δ — некоторая известная константа, приравняем в выражении (15.7.1) ро к *8*. Тогда

нулевая гипотеза будет отклонена, если значение критериальной статистики попадет в верхнюю или нижнюю 2,5 %-ную область t_{n-2}-распределения.

Для проверки гипотезы $\sigma^2 = \sigma^2_0$, где σ^2_0 — известная константа, приравняем в выражении (15.7.3) σ^2 к σ^2_0. Нулевая гипотеза будет отклонена, если значение критериальной статистики попадет в верхнюю или нижнюю 2,5 %-ную область распределения χ^2_{n-2}.

Пример 15.8.1. Проверим нулевую гипотезу о том, что данные из примера 15.7.1 получены из совокупности, регрессия которой имеет коэффициент наклона $\beta_1 = 3$.

Первый способ проверки — использование (15.7.2) и вычисление критериальной статистики.

$$t_{18} = (2{,}972\,93 - 3)/[11{,}406/\{2870 - 20 \times (10{,}5)^2\}]^{1/2} = -0{,}2067.$$

Это значение не является значимым при 5 %-ном уровне, поэтому у нас нет причин отклонить нулевую гипотезу.

Таблица 15.8.1. Значения r, полученные вычитанием $3x$ из y

x	r	x	r	x	r	x	z
1	14	6	10	11	11	16	12
2	7	7	5	12	11	17	7
3	13	8	4	13	6	18	7
4	8	9	7	14	12	19	7
5	5	10	16	15	10	20	10

Другой способ — вычитание $3x_i$ из y_i (см. табл. 15.7.1), получение значения $\{z_i\}$ в табл. 15.8.1 и построение таблицы дисперсионного анализа (см. табл. 15.8.2). Величина

$$F_{1,18} = 0{,}487/11{,}406 = 0{,}0427$$

не значима при 5 %-ном уровне, поэтому у нас нет причин отклонять нулевую гипотезу.

Таблица 15.8.2. Дисперсионный анализ данных из примера 15.8.1

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	0,487	1	0,487
Остаток	205,313	18	11,406
Общая вариация	205,800	19	—

Напомним читателю, что F-статистика является квадратом t-статистики, а суммы квадратов отклонений в табл. 15.7.2 и 15.8.2 совпадают, поэтому предложенные два критерия математически

эквивалентны. Разность между оценками коэффициента наклона в табл. 15.7.1 и 15.8.1 равна в точности 3,0.

Литература: [19, с. 18—21; русский перевод с. 28—29], [54, с. 362—363], [105, с. 209—210].

15.9. МАТРИЧНЫЙ ПОДХОД В РЕГРЕССИОННОМ АНАЛИЗЕ

В параграфе 15.3 было показано, как при применении метода наименьших квадратов численные процедуры могут быть записаны с помощью теории матриц. Матричный метод широко используется в статистическом анализе. Большинство необходимых векторов и матриц было введено в параграфе 15,3. Определим двумерный вектор-столбец p с элементами β_0 и β_1 и n-мерный вектор-столбец с элементами $\{e_i\}$ $(i = 1, 2, \ldots, n)$.

Статистическая модель (15.4.1) может быть записана в следующем виде:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{15.9.1}$$

Обозначим матрицу, обратную к S, через $\mathbf{S}^{-1}$. Тогда нетрудно показать с помощью алгебраических дополнений (см. параграф 1.10), что

$$\mathbf{S}^{-1} = \begin{pmatrix} S_{00}^{-1} & S_{01}^{-1} \\ S_{10}^{-1} & S_{11}^{-1} \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ -\sum x_i & n \end{pmatrix} \Big/ (n \sum x_i^2 - n^2 \bar{x}^2). \tag{15.9.2}$$

Матрица ковариаций b_0 и b_1 есть $\sigma^2\mathbf{S}^{-1}$, поэтому

$$\operatorname{var} b_0 = \sigma^2 S_{00}^{-1}, \tag{15.9.3}$$

$$\operatorname{var} b_1 = \sigma^2 S_{11}^{-1}, \tag{15.9.4}$$

$$\operatorname{cov}(b_0, b_1) = \sigma^2 S_{01}^{-1}. \tag{15.9.5}$$

Легко видеть, что t-статистики (15.7.1) и (15.7.2) при построении доверительных интервалов для β_0 и β_1 могут быть записаны в виде

$$t_{n-2} = (b_0 - \beta_0)/(s^2 S_{00}^{-1})^{\frac{1}{2}}, \tag{15.9.6}$$

$$t_{n-2} = (b_1 - \beta_1)/(s^2 S_{11}^{-1})^{\frac{1}{2}}. \tag{15.9.7}$$

Суммы квадратов в таблице дисперсионного анализа 15.5.1 могут быть представлены несколькими способами:

$$\text{с. к. регрессии} = \mathbf{b'X'y} - n\bar{y}^2, \tag{15.9.8}$$

$$\text{с. к. регрессии} = \mathbf{b'Sb} - n\bar{y}^2, \tag{15.9.9}$$

$$\text{общая с. к.} - \mathbf{y'y} - n\bar{y}^2. \tag{15.9.10}$$

Хотя запись суммы квадратов регрессии в табл. 15.5.1 не совпадает с (15.9.8) и (15.9.9), в действительности эта сумма равна им.

Все результаты обобщаются на случай криволинейных и множественных регрессий.

Литература: [19, с. 53—56; русский перевод с. 53—56], [54, с. 354—355].

15.10. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ ДЛЯ СРЕДНЕГО ЗАВИСИМОЙ ПЕРЕМЕННОЙ ПРИ ЗАДАННОМ ЗНАЧЕНИИ НЕЗАВИСИМОЙ ПЕРЕМЕННОЙ

Предположим, что, построив модель линейной регрессии по n наблюдениям, мы хотим оценить ожидаемое значение, отвечающее значению $X = x_0$. Точечная оценка ожидаемого значения y есть

$$y_0 = b_0 + b_1 x_0. \qquad (15.10.1)$$

Определим вектор-столбец $\mathbf{x}_0 = (1, x_0)$, тогда (15.10.1) можно записать как

$$y_0 = \mathbf{x}_0 \mathbf{b}. \qquad (15.10.2)$$

$\{b_j\}$ являются точечными оценками $\{\beta_j\}$. Для построения 95 %-ного доверительного интервала для $\hat{y}$, ожидаемого значения y в точке x_0, приравняем

$$t_{n-2} = (Y - y_0)/\{s^2(\mathbf{x}_0 \mathbf{S}^{-1} \mathbf{x}_0')\}^{\frac{1}{2}} \qquad (15.10.3)$$

к верхней и нижней 2,5 %-ным точкам t_{n-2}-распределения. Вероятно, излишне предостерегать читателя об опасности бездумного экстраполирования по подобранной регрессии, как бы хорошо она ни была подогнана к имеющимся точкам. Экстраполяция не имеет, вообще говоря, теоретического обоснования. Результаты данного параграфа могут быть обобщены для криволинейной и множественной регрессий.

Пример 15.10.1. Найти 95 %-ный доверительный интервал для среднего значения y в точке $x = 10$ по данным табл. 15.7.1.

Как следует из примера 15.7.1, $b_0 = 9{,}384\,21$ и $b_1 = 2{,}972\,93$. Отсюда точечная оценка среднего y равна:

$$y = 9{,}384\,21 + 10 \times 2{,}972\,93 = 39{,}1135.$$

Средний квадрат отклонений s^2 есть 11,406, а обратная матрица имеет вид:

$$\mathbf{S}^{-1} = \begin{pmatrix} 20 & 210 \\ 210 & 2870 \end{pmatrix}^{-} = \begin{pmatrix} 0{,}21578947 & -0{,}01578947 \\ -0{,}015\,789\,47 & 0{,}001\,503\,75 \end{pmatrix}.$$

Тогда

$$x_0 = (1\ \ 10); \quad \mathbf{S}^{-1}\mathbf{x}_0' = \begin{pmatrix} 0{,}057\,894\,77 \\ -0{,}000\,751\ \ 97 \end{pmatrix} \text{ и } \mathbf{x}_0 \mathbf{S}^{-1} \mathbf{x}_0' = 0{,}050\,375.$$

Для вычисления 95 %-ных доверительных границ для среднего $\hat{y}$ решим уравнения

$$(y - 39{,}1135)/\sqrt{11{,}406 \times 0{,}050\,375} = \pm 2{,}101,$$

которые приводят нас к $39{,}11 \pm 1{,}59$.

Литература: [19, с. 21—24; русский перевод с. 32—33], [54, с. 363—365], [105, с. 210—213].

15.11. ПРАВИЛЬНУЮ ЛИ МОДЕЛЬ МЫ ВЫБРАЛИ?

После того как линия регрессии построена, естественно, возникает вопрос: а правильную ли модель мы выбрали? Методы, которые будут сейчас описаны, с одинаковым успехом можно применять как для парной линейной, так и для криволинейной и множественной регрессии.

Когда мы формулировали модель (15.4.1), предполагалось, что ошибки $\{e_i\}$ являются независимыми и распределенными по нормальному закону с нулевым математическим ожиданием и дисперсией σ^2. Разности между наблюдаемыми и расчетными значениями y, т. е. *остатки* (отклонения), в принципе также должны

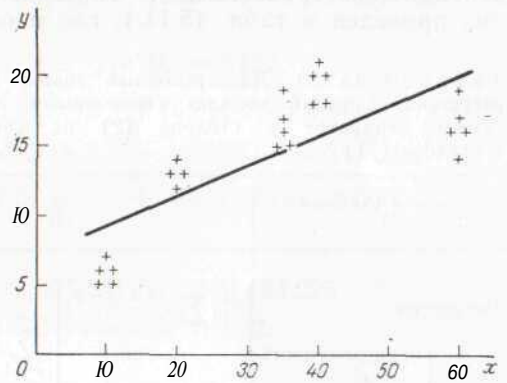


Рис. 15.11.1. Пример неадекватной подгонки

удовлетворять этим предположениям. В противном случае у нас могут возникнуть серьезные опасения относительно правильности выбранной модели.

Необходимо выполнить следующую проверку:

1. Убедиться в том, похожа ли совокупность отклонений на выборку из нормального распределения с нулевым математическим ожиданием.

2. Последовательно нанести отклонения на график и посмотреть, можно ли оправдать предположение о постоянстве дисперсии или же отклонения на одном из концов имеют большую дисперсию?

3. Выяснить, можно ли из графика отклонений сделать вывод о несоответствии модели линейной регрессии? (Как видно из рис. 15.11.1, квадратичная кривая, например, лучше подходила бы к данным, чем прямая линия.)

4. Тщательно исследовать выбросы *. Выбросы должны быть удалены из совокупности данных, если они представляют собой очевидную ошибку эксперимента.

Если для каждого x имеется несколько повторных наблюдений $\{y_i\}$, то неадекватность подгонки может быть проверена статистически. Предположим, имеется n_i наблюдений в точке x_i, которые обозначим $y_{i1}, \ldots, y_{in_i}$. Если линия регрессии адекватна данным, среднее n_i наблюдений $\bar{y}_i$ будет лежать близко к расчетному значению Y_i и величина $(\bar{y}_{i\cdot} - Y_i)^2$ будет мала. Если аппроксимация плохая, значение последнего квадрата будет велико. Мерой *неадекватности,* таким образом, может служить величина

$$\sum_i n_i (\bar{y}_{i\cdot} - Y_i)^2.$$

При повторных наблюдениях сумма квадратов отклонений в табл. 15.5.1 в свою очередь может быть разбита на две положительные части: только что приведенная сумма квадратов как мера неадекватности и *сумма квадратов чистой ошибки.* Последняя представляет собой сумму квадратов отклонений значений $\{y_i\}$ относительно среднего $\{\bar{y}_{i\cdot}\}$. Соответствующий дисперсионный анализ приведен в табл. 15.11.1, где p обозначает число различных

Таблица 15.11.1. **Дисперсионный анализ неадекватности. Парная линейная регрессия. Средний квадрат (отмеченный звездочками) получается делением суммы квадратов в столбце (2) на соответствующую степень свободы в столбце (3)**

Источник вариации (1)	с. к. (2)	с. с. (3)	ср. к. (4)=(2)/(3)
Регрессия	$b_1^2(\sum x^2 - n\bar{x}^2)$	1	**
Остаток: неадекватность	$\sum_{i=1}^{p} (T_{i\cdot}^2/n_i) - T_{\cdot\cdot}^2/n$ — с. к. регрессии	$p-2$	*
Остаток: чистая ошибка	находится как разность	$n-p$	*
Общая вариация	$\sum y^2 - n\bar{y}^2$	$n-1$	—

значений *XL* Проверка неадекватности достигается подсчетом величины

$$F_{p-2,\ n-p} = \frac{\text{средний квадрат неадекватности}}{\text{средний квадрат чистой ошибки}}. \qquad (15.11.1)$$

* Т. е. наблюдения, отклонения для которых резко отличаются от остальной совокупности отклонений.— *Примеч. пер.*

Для проверки адекватности подгонки линейной регрессии применяется обычный $F_{1,\ n-2}$-критерий. Вычисления несколько утомительны. Обозначим сумму значений $\{y_{ij}\}$ в точке x_i через $T_{i\cdot}$, общую сумму $\{y_{ij}\}$ — через $T_{\cdot\cdot}$. Суммы квадратов находят следующим образом:

1) вычисляют сумму квадратов регрессии (см. табл. 15.5.1);
2) вычисляют общую сумму квадратов (см. табл. 15.5.1);
3) вычисляют остаточную сумму квадратов как разность: (2)—(1);
4) вычисляют [7] сумму регрессии и сумму квадратов неадекватности по формуле

$$(T_{1\cdot}^2/n_1 + \ldots + T_{p\cdot}^2/n_p) - T_{\cdot\cdot}^2/n; \qquad (15.11.2)$$

5) вычисляют сумму квадратов неадекватности как разность: (4)—(1);
6) вычисляют сумму квадратов чистой ошибки как разность: (3)—(5).

Пример 15.11.1. С помощью метода наименьших квадратов подберем прямую линию для данных из табл. 15.11.2 и проверим неадекватность регрессии. Начнем с вычисления следующих величин:

$$\sum\sum y^2 = 5^2 + 6^2 + \ldots + 16^2 = 5452,$$
$$\sum x = (10 \times 5) + \ldots + (60 \times 5) = 805,$$
$$\sum x^2 = (10^2 \times 5) + \ldots + (60^2 \times 5) = 34\,225;$$
$$\text{S Z } xy = (10 \times 5) + \ldots + (60 \times 16) = 13\,000.$$

По формуле (15.2.6)

$$b_1 = \{13\,000 - 24 \times (805/24)\,(342/24)\}/\{34\,225 - 24 \times (805/24)^2\} = 0{,}2\,116\,222.$$

Из табл. 15.5.1 находим:

с. к. регрессии = $(0{,}2\,116\,222)^2\{34\,225 - 24 \times (805/24)^2\} = 323{,}52$,

общая с. к. = $5452 - 24 \times (342/24)^2 = 578{,}50$,

остаточная с. к. = $578{,}50 - 323{,}52 = 254{,}98$.

По формуле (15.11.2)

с. к. регрессии + с. к. неадекватности = $(29^2/5) + (52^2/4) + \ldots$
$\ldots + (82^2/5) - (342^2/24) = 542{,}10$.

Вычитаем:

с. к. неадекватности = $542{,}10 - 323{,}52 = 218{,}58$,

с. к. чистой ошибки = $254{,}98 - 218{,}58 = 36{,}40$.

[7] Вычисление суммы регрессии и суммы квадратов неадекватности основано на методе однофакторного дисперсионного анализа.

Таблица **15.11.2. Данные для подбора регрессии из примера 15.11.1 (24 наблюдения взаимно независимы)**

x_i	y_{ij}					n_i	T_i.
10	5	6	5	6	7	5	29
20	12	13	14	13		4	52
35	17	19	16	15	15	5	82
40	18	20	21	18	20	5	97
60	17	19	16	14	16	5	82
Всего						24	342

Таблица **15.11.3. Дисперсионный анализ неадекватности. Числовой пример**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	323,52	1	323,52
Остаток: неадекватность	218,58	3	72,86
Остаток: чистая ошибка	35,40	19	1,92
Остаток	254,98	22	11,59
Общая вариация	578,50	23	—

Эти значения собраны в табл. 15.11.3, где показаны также средние квадраты. Для проверки на неадекватность вычислим $F_{3,19} = 72{,}86/1{,}92 = 37{,}95$. Это значение имеет высокую значимость. Отсюда следует, что полученная прямая плохо подходит к данным. Неадекватность также легко обнаружить на графике, содержащем точки наблюдений и линейную регрессию (см, рис. 15.11.1).

Литература: [7, с. 366—375; русский перевод с. 351—362], [19, с. 26—32, 86—100; русский перевод с. 33—40], [105, с. 215—223].

15.12. НЕРАВНЫЕ ДИСПЕРСИИ. ВЗВЕШЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Результаты эксперимента не всегда имеют одинаковую надежность и исследователь часто может что-либо утверждать об относительной надежности своих наблюдений. Иногда отсутствие гомогенности дисперсий может быть не выяснено вплоть до построения регрессии и нанесения на график ее отклонений. В других случаях предположение о постоянстве дисперсий ошибок $\{e_i\}$ в (15.4.1) не выполняется с достаточной очевидностью. Даже при установленном отсутствии гомогенности дисперсий (например, некоторые дисперсии в 10 раз больше других) дополнительные вычисления с применением взвешенного метода наименьших квадратов не являются строго необходимыми — оценки b_0 и b_1 по обычному методу наименьших квадратов по-прежнему остаются несмещенными оценками β_0 и β_1, хотя они и не представляют собой оценки с минимальной дисперсией (см. параграф 13.1). Если неравенство дисперсий замечено, то следует применять взвешенный метод наименьшихквадратов.

Рассмотрим случай n точек и предположим, что дисперсия в i-й точке равна σ^2/w_i. Величины $\{w_i\}$ называются *весами*.

Если нас интересуют только оценки b_0 и b_1 и нашей целью не является выполнение других статистических вычислений, то взвешенный метод наименьших квадратов применить очень просто: мы считаем[8], что y_i наблюдается целое число раз, равное Nw_i, где N — достаточно большое положительное целое число. Тогда стандартные процедуры метода наименьших квадратов приведут нас к искомым значениям b_0 и b_1.

Метод, рассмотренный в предыдущем параграфе, не подходит для статистических целей, поэтому необходим другой подход. Его лучше описать в терминах матричной теории. Предположим, что $n = 4$. Определим

$$\mathbf{X}_w = \begin{pmatrix} \sqrt{w_1} & x_1\sqrt{w_1} \\ \sqrt{w_2} & x_2\sqrt{w_2} \\ \sqrt{w_3} & x_3\sqrt{w_3} \\ \sqrt{w_4} & x_4\sqrt{w_4} \end{pmatrix}, \quad \mathbf{y}_w = \begin{pmatrix} y_1\sqrt{w_1} \\ y_2\sqrt{w_2} \\ y_3\sqrt{w_3} \\ y_4\sqrt{w_4} \end{pmatrix}, \quad \mathbf{S}_w = \mathbf{X}_w'\mathbf{X}_w.$$

Записывая $\mathbf{X}_w$ вместо X, $\mathbf{y}_w$ вместо y, $\mathbf{S}_w$ вместо S и подставляя вместо элементов $\mathbf{S}^{-1}$ соответствующие элементы $\mathbf{S}^{-1}$ и вместо y

$$y_w = (y_1w_1 + y_2w_2 + \ldots + y_nw_n)/(w_1 + w_2 + \ldots + w_n),$$

применяем стандартные формулы регрессионного анализа (153 1), (15.3.3), (15.3.4), (15.5.3), (15.7.3), (15.9.1), (15.9.3), (15.94), (15.9.5), (15.9.6), (15.9.7), (15.10.2) и (15.10.3). Если в формулах (15.9.8), (15.9.9) и (15.9.10) вместо n подставить $\sum w_i$, ими можно пользоваться, как и ранее. Перечисленные формулы дают все необходимое.

Пример 15.12.1. Эксперимент привел к следующим результатам:

$$(x_1, y_1) = (1, 7),$$
$$(x_2, y_2) = (3, 18),$$
$$(x_3, y_3) = (4, 30),$$
$$(x_4, y_4) = (5, 35).$$

Известно, что стандартные отклонения наблюдений для x_1 и x_4 приблизительно совпадают, а для наблюдений x_2 и x_3 они примерно в 3 раза больше. Найдем линейную регрессию с помощью взвешенного метода наименьших квадратов.

[8] Это представление будет очевидным, если вспомнить, что выборочное среднее Nw_i наблюдений из совокупности с дисперсией σ^2 имеет дисперсию $\sigma^2/(Nw_i)$.

В данном примере точность наблюдений для x_1 и x_4 примерно в 3 раза больше (в смысле стандартного отклонения), чем точность x_2 и x_3. Поэтому полагаем, что $w_1 = w_4 = 9$, а $w_2 = w_3 = 1$. Для вычисления b_0 и b_1 представим себе, что наблюдения x_1 и x_4 повторены 9 раз. Таким образом, $n = 20$ и

$$\sum x_i = 9 \times 1 + 3 + 4 + 9 \times 5 = 61,$$
$$\sum x_i^2 = 9 \times 1^2 + 3^2 + 4^2 + 9 \times 5^2 = 259,$$
$$\sum y_i = 9 \times 7 + 18 + 30 + 9 \times 35 = 426,$$
$$\sum x_i y_i = 9 \times 1 \times 7 + 3 \times 18 + 4 \times 30 + 9 \times 5 \times 35 = 1812.$$

Из формул (15.2.6) получаем

$$b_1 = \{1812 - 20 \times (61/20)\,(426/20)\}/\{259 - 20 \times (61/20)^2\} = 7{,}0281,$$
$$b_0 = (426/20) - 7{,}0281 \times (61/20) = -0{,}1357.$$

С помощью (15.2.1) найдем расчетные значения: 6,892; 20,949; 27,977 и 35,005. Отметим, что прямая линия прошла близко к точкам x_1 И x_4.

Пример 15.12.2. Найдем линейную регрессию, используя взвешенный метод наименьших квадратов, для данных из примера 15.12.1.

Вычислим доверительные интервалы для β_0 и β_1. Определим

$$\mathbf{X}_w = \begin{pmatrix} 3 & 3 \\ 1 & 3 \\ 1 & 4 \\ 3 & 15 \end{pmatrix}, \quad \mathbf{y}_w = \begin{pmatrix} 21 \\ 18 \\ 30 \\ 105 \end{pmatrix}, \quad \mathbf{S}_w = \begin{pmatrix} 20 & 61 \\ 61 & 259 \end{pmatrix}.$$

Нам также необходимо найти

$$y_w = (7 \times 9 + 18 \times 1 + 30 \times 1 + 35 \times 9)/20 = 21{,}3,$$

$$\mathbf{S}_w^{-1} = \begin{pmatrix} 0{,}177\,518\,848 & -0{,}041\,809\,458 \\ -0{,}041\,809\,458 & 0{,}013\,708\,019 \end{pmatrix}, \quad \mathbf{X}_w \mathbf{y}_w = \begin{pmatrix} 426 \\ 1812 \end{pmatrix}.$$

Из формулы (15.3.4) получаем

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \mathbf{S}_w^{-1} \mathbf{X}_w \mathbf{y}_w = \begin{pmatrix} -0{,}1357 \\ 7{,}0281 \end{pmatrix},$$

что совпадает с ранее полученным результатом.

Применяя формулы (15.9.8) и (15.9.10) с индексом w и подставляя $\sum w_i$ вместо n, заполним клетки таблицы дисперсионного анализа 15.12.1. Вычислим

$$F_{1,2} = 3603{,}31/6{,}45 = 559.$$

Этот результат весьма значим, поэтому вычисление доверительных интервалов для β_0 и β_1 имеет большое значение.

Таблица 15.12.1. Дисперсионный анализ. Взвешенная регрессия

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	3603,31	1	3603,31
Остаток	12,89	2	6,45
Общая вариация	3616,20	3	—

Для нахождения 95 %-ных доверительных интервалов для β_0 и β_1 приравняем (15.9.6) и (15.9.7) к верхней и нижней 2,5 %-ным точкам t_2-распределения:

$$(-0{,}1357 - \beta_0)/\sqrt{6{,}45} \times 0{,}177\,518\,848 = \pm 4{,}303,$$
$$(7{,}0281 - \beta_1)/\sqrt{6{,}45} \times 0{,}013\,708\,019 = \pm 4{,}303.$$

Доверительными интервалами будут $-0{,}1357 + 4{,}6044$ и $7{,}0281 \pm 1{,}2795$. Границы весьма широки, это следствие малого объема выборки.

Литература: [19, с. 77—80; русский перевод с. 86—89].

15.13. ОБЩИЕ ЧЕРТЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА И ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Двумерное нормальное распределение было введено в параграфе 9.6. Критерий проверки значимости коэффициента корреляции, основанный на этом распределении, был описан в параграфах 12.31—12.33. Читатель мог заметить некоторое сходство между методами корреляции двумерного нормального распределения и парной линейной регрессии: например, между линией регрессии, построенной по методу наименьших квадратов (15.2.7), и уравнением условного математического ожидания (9.6.2); критериальной статистикой коэффициента корреляции **(12.31.2)** и регрессии (15.5.4).

Оба метода могут применяться, только если предполагаемая связь между двумя переменными линейна. Регрессионная техника, однако, может быть обобщена и на кривые (см. гл. 16). В корреляционном анализе x и y — случайные переменные, однако здесь необходимо также строгое предположение о двумерной нормальности. При этом нас интересует мера зависимости в линейной связи, а не сами параметры этой зависимости. В регрессионном анализе мы различаем независимую переменную, которая по предположению свободна от ошибок, и зависимую переменную y, условное распределение которой при заданном x является нормальным. Теперь наша цель — как можно больше узнать о параметрах регрессии. Важно отметить, что для нормально распределенных переменных x и y распределение y при заданном x нормально с постоянной дисперсией, не зависящей от x, поэтому методы регрессии

могут применяться и для решения задач с данными, распределенными по двумерному нормальному закону.

Техника и в том, и в другом случае может быть обобщена для ситуации с тремя или большим числом переменных. Обобщения парной линейной регрессии приводятся в гл. 17.

Литература: [8, с. 281—282], [50, с. 435—436; русский перевод с. 472—474], [105, с. 198—199].

15.14. УПРАЖНЕНИЯ

1. С помощью метода наименьших квадратов постройте прямую линию по данным табл. 15.14.1.

Таблица 15.14.1. 36 независимых результатов эксперимента

x	Наблюдения y			
1	123	119	125	128
14	150	142	173	158
21	191	182	175	179
28	206	229	214	218
35	251	245	241	252
42	258	266	276	274
49	307	302	295	301
56	321	328	326	330
63	353	342	346	351

2. Постройте для данных табл. 15.14.1 матрицу S и вектор X'y.
3. Вычислите коэффициент корреляции r для данных табл. 15.14.1 и с помощью формулы (15.5.4) проверьте гипотезу о значимости линейной регрессии.
4. Постройте таблицу дисперсионного анализа для регрессии, найденной по данным табл. 15.14.1. Проверьте значимость линейной регрессии и установите численную зависимость между t и F.
5. Найдите 95 %-ный доверительный интервал для β_0, β_1 и σ^2.
6. Выбрана ли для данных табл. 15.14.1 правильная модель?
7. Примените невзвешенный метод наименьших квадратов к данным табл. 15.12.1. Сравните ваши результаты с представленными в книге.

16. КРИВОЛИНЕЙНАЯ РЕГРЕССИЯ

В этой главе рассматриваются криволинейные регрессии — обобщение парных линейных регрессий. Общая процедура метода наименьших квадратов описывается в параграфах 16.1 и 16.2. В параграфах 16.3 и 16.4 обсуждаются статистические модели и критерии проверки значимости криволинейных регрессий. Процедуры оценивания и проверки гипотез разбираются в параграфах 16.5—16.8. В параграфах 16.9 и 16.10 рассматриваются вопросы неадекватности и взвешенный метод наименьших квадратов. Проблемы линейной регрессии, график которой проходит через начало координат или другую фиксированную точку на плоскости, описывается в параграфе 16.11. Глава кончается двумя параграфами, посвященными ортогональным полиномам.

16.1. ВВЕДЕНИЕ. НЕСКОЛЬКО ПРОСТЫХ ПРИМЕРОВ КРИВОЛИНЕЙНЫХ РЕГРЕССИЙ

Методы гл. 15 можно без труда распространить на криволинейные регрессии. Матричное обозначение позволяет оперировать уравнениями и системами уравнений в общем виде, что существенно облегчает работу. Мы советуем читателю обратиться к соответствующим параграфам предыдущей главы (см. параграф 15.3 и в особенности параграф 15.9), где вводятся основные определения и обозначения матричной алгебры.

В уравнение прямой линии x входит с нулевым и единичным показателем степени, а при подгонке прямой по методу наименьших квадратов мы использовали матрицу X с двумя столбцами, один из которых соответствует нулевой, а другой — первой степени $\{x_i\}$. Полином второй степени

$$Y = b_0 + b_1 x + b_2 x^2 \tag{16.1.1}$$

содержит нулевую, первую и вторую степени x, и для подгонки этого полинома по методу наименьших квадратов к n точкам необходимо применить матрицу X с тремя столбцами. Первый, второй и третий столбцы соответствуют нулевой, первой и второй степенями значений $\{x_i\}$. Таким образом,

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_n & x_n \end{pmatrix}. \tag{16.1.2}$$

Наша задача — найти три константы b_0, b_1 и b_2, которые образуют трехмерный вектор-столбец b. Как и прежде, наблюдения $\{y_i\}$ образуют вектор-столбец y, а выравненные значения $\{Y_i\}$ — вектор-столбец Y. Оба эти вектора содержат n элементов. Легко видеть, что

$$\mathbf{S} = \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix} \text{ и } \mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{pmatrix}. \tag{16.1.3}$$

Оценки по методу наименьших квадратов b_0, b_1 и b_2 находятся решением матричного уравнения $\mathbf{Sb} = \mathbf{X}'\mathbf{y}$ (см. параграф 15.3).

Экспоненциальная кривая [1]

$$Y = b_0 + b_1 e^x \tag{16.1.4}$$

[1] Экспоненциальные кривые вида $Y = Ae^{Bx}$, где Л и В неизвестны сводят обычно к линейной регрессии логарифмированием (см. параграф 18.1).

содержит нулевую степень x и e^x, и для подгонки этой кривой к n точкам необходимо построить матрицу X с двумя столбцами. Первый столбец содержит нулевые степени $\{x_i\}$, а второй — значения $\{\exp(x_i)\}$. Итак,

$$\mathbf{X} = \begin{pmatrix} 1 & \exp(x_1) \\ 1 & \exp(x_2) \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & \exp(x_n) \end{pmatrix}. \qquad (16.1.5)$$

Необходимо найти константы b_0 и b_1, которые образуют двумерный вектор b. Наблюдения $\{y_i\}$ и расчетные значения $\{Y_i\}$ образуют соответственно вектор-столбцы y и Y. Легко видеть, что

$$\mathbf{S} = \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum \exp(x_i) \\ \sum \exp(x_i) & \sum \{\exp(x_i)\}^2 \end{pmatrix} \text{ и } \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum y_i \\ \sum \{\exp(x_i)\} y_i \end{pmatrix}. \qquad (16.1.6)$$

Оценки по методу наименьших квадратов находятся решением матричного уравнения Sb = X'y (см. параграф 15.3).

Пример 16.1.1. С помощью метода наименьших квадратов подберем квадратичную кривую (16.1.1) по четырем точкам:

$$(x_1, y_1) = (1,5),$$
$$(x_2, y_2) = (3,40),$$
$$(x_3, y_3) = (4,109),$$
$$(x_4, y_4) = (5,297).$$

Как следует из (16.1.3),

$$\mathbf{S} = \begin{pmatrix} 4 & 13 & 51 \\ 13 & 51 & 217 \\ 51 & 217 & 963 \end{pmatrix} \text{ и } \mathbf{X}'\mathbf{y} = \begin{pmatrix} 451 \\ 2046 \\ 9534 \end{pmatrix}.$$

Нормальные уравнения (в матричном виде Sb = X'y) могут быть записаны как

$$4b_0 + 13b_1 + 51b_2 = 451,$$
$$13b_0 + 51b_1 + 217b_2 = 2046,$$
$$51b_0 + 217b_1 + 963b_2 = 9534.$$

Находим, что $b_0 = 89{,}79988$, $b_1 = -113{,}12719$ и $b_2 = 30{,}636\,352$, поэтому расчетные значения y равны соответственно 7,309; 26,145; 127,473; 290,073. Квадратичная кривая и точки наблюдений показаны на рис. 16.1.1.

Пример 16.1.2. По методу наименьших квадратов подберем экспоненциальную кривую (16.1.4) к четырем точкам из примера 16.1.1. Как следует из (16.1.6),

$$\mathbf{S} = \begin{pmatrix} 4 & 225{,}8\ 151\ 278 \\ 225{,}8\ 151\ 278 & 25\ 418{,}241\ 62 \end{pmatrix} \text{ и } \mathbf{X}\mathbf{y} = \begin{pmatrix} 415 \\ 50856{,}91\ 948 \end{pmatrix}.$$

Нормальные уравнения (в матричном виде Sb = X'y) могут быть записаны в виде

$$4b_0 + 225{,}8\ 151\ 278b_1 = 451,$$
$$225{,}8\ 151\ 278b_0 + 25\ 418{,}24\ 162b_1 = 50846{,}91\ 948.$$

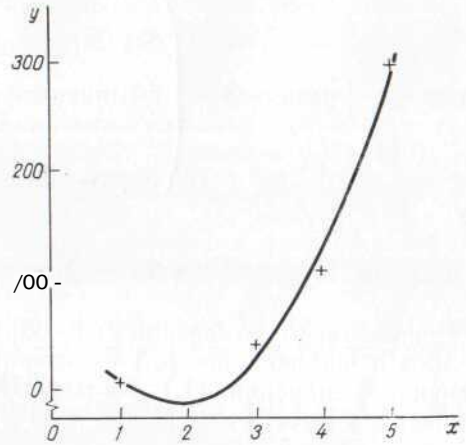


Рис. 16.1.1. Квадратичная кривая, построенная по методу наименьших квадратов. Точки наблюдений отмечены крестиками

Находим $b_0 = -0{,}362\,598\,010$ и $b_1 = 2{,}003631\,893$. Расчетные значения равны 5,084; 39,881; 109,032 и 297,003. Подгонка оказалась хорошей.

Литература: [7, с. 447; русский перевод с. 362—363], [16, с. 243—271], [46, с. 177—180], [91, с. 447—472; русский перевод с. 417—437], [102, с. 273—275].

16.2. ОБОБЩЕННЫЙ КРИВОЛИНЕЙНЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Матричные методы, описанные в параграфе 16.1, могут быть применены к любой функциональной форме, линейной по коэффициентам $\{b_i\}$. Например, матричный подход может быть применен к кривой вида

$$Y = b_0 + b_1 \sin(\pi e^x) + b_2 \ln x,$$

однако его нельзя применять к кривой с уравнением [2]

$$Y = b_0 + b_1 \exp(b_2 x).$$

[2] Здесь необходимо воспользоваться нелинейными методами, рассмотренными в гл. 18 (см. также сноску 1).

В общем случае необходимо определить $k+l$ коэффициентов кривой вида

$$V = b_0 g_0(x) + b_1 g_1(x) + \ldots + b_k g_k(x). \qquad (16.2.1)$$

Функции $\{g_i(x)\}$ не должны содержать неизвестных параметров, подлежащих оцениванию.

Для подгонки кривой (16.2.1) к точкам по методу наименьших квадратов необходимо построить матрицу с $k+l$ столбцами. Первый столбец составлен из значений $\{g_0(x_i)\}$, второй — из значений $\{g_1(x_i)\}$, последний — из значений $\{g_k(x_i)\}$. Нетрудно проверить, что матрица S = X'X равна:

$$\mathbf{S} = \begin{pmatrix} S_{00} & S_{01} & S_{02} & \ldots & S_{0k} \\ S_{10} & S_{11} & S_{12} & \ldots & S_{1k} \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ S_{k0} & S_{k1} & S_{k2} & \ldots & S_{kk} \end{pmatrix}, \qquad (16.2.2)$$

где

$$S_{rs} = S_{sr} = \sum_{i=1}^{n} g_r(x_i)\, g_s(x_i). \qquad (16.2.3)$$

Неизвестные коэффициенты образуют $k+1$-мерный вектор-столбец b, наблюдения $\{y_i\}$ — n-мерный вектор-столбец y, а n выравненных значений $\{Y_i\}$ — n-мерный вектор-столбец Y. Вектор-столбец $\mathbf{X'y}$ имеет размерность $k+1$ и равен:

$$\mathbf{X'Y} = \begin{pmatrix} \sum_i y_i g_0(x_i) \\ \sum_i y_i g_1(X_i) \\ \cdot \\ \cdot \\ \cdot \\ \sum_i y_i g_k(x_i) \end{pmatrix}. \qquad (16.2.4)$$

Для нахождения $k+l$ коэффициентов $\{b_i\}$ решим $k+l$ линейных уравнений Sb = X'y.

Пример 16.2.1. Для квадратичной кривой (16.1.1) $g_0(x) = 1$, $g_1(x) = x$ и $g_2(x) — x^2$ Матрица S и вектор X'y задаются выражением (16.1.3).

Пример 16.2.2. Для экспоненциальной кривой (16.1.4) $g_0(x) = 1$ и $g_1(x) = e^x$. Матрица S и вектор X'y задаются выражением (16.1.6).

Пример 16.2.3. Методом наименьших квадратов подобрать кривую вида $Y — b_1 e^x$ к четырем точкам из примера 16.1.1.

В данном случае необходимо найти только одну константу, поэтому матрица X имеет только один столбец, составленный из значений $\{\exp x_i\}$. Матрица S = X'X размерности 1x1 с элементом

$$\sum_{i=1}^{4} \{\exp x_i\}^2 = 25\,418{,}24\ 162.$$

Вектор-столбец X'y содержит один элемент

$$\sum_{i=1}^{4} \{\exp x_i\}\, y_i = 50\,846{,}91\,948.$$

Нормальное уравнение Sb = X'y имеет простой вид:

$$25{,}418{,}24\ 162 b_1 = 50\,846{,}91\,948,$$

откуда $b_1 = 2{,}000410580$. Выравненными значениями являются 5,438; 40,179; 109,219 и 296,887 соответственно.

Литература: [19, с. 58—66; русский перевод с. 137—143], [50, с. 423—434; русский перевод с. 487—498], [91, с. 447—472; русский перевод с. 417—437], [105, с. 268—270].

16.3. КРИВОЛИНЕЙНАЯ РЕГРЕССИЯ. СТАТИСТИЧЕСКАЯ МОДЕЛЬ

В параграфах 16.1 и 16.2 метод наименьших квадратов был описан как вычислительная процедура. Сформулируем теперь соответствующую статистическую модель.

При построении квадратичной зависимости (16.1.1) мы неявно предполагали следующую статистическую модель:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \qquad (16.3.1)$$

где $\{x_i\}$ по условию свободны от ошибок. Величины $\{\beta_i\}$ неизвестны, а $\{e_i\}$ — случайные независимые переменные, распределенные по нормальному закону с нулевым математическим ожиданием и одинаковой дисперсией. Оценками для β_0, β_1 и β_2 являются *bo, b* и b_2 соответственно. Для экспоненциальной кривой (16.1.4) предполагается, что статистической моделью, лежащей в ее основе, является

$$y_i = \beta_0 + \beta_1 \exp x_i + e_i. \qquad (16.3.2)$$

Для криволинейной зависимости общего вида (16.2.1) такой статистической моделью будет:

$$y_i = \beta_0 g_0(x_i) + \ldots + \beta_k g_k(x_i) + e_i. \qquad (16.3.3)$$

В уравнениях (16.3.2) и (16.3.3) $\{x_i\}$ свободны от ошибок. Все эти модели можно представить в матричном виде (15.9.1), при этом совокупности {p,:} и $\{e_i\}$ составляют соответственно вектор-столбцы β и e.

Литература: [19, с. 58—66; русский перевод с. 137—143].

16.4. ПРОВЕРКА ЗНАЧИМОСТИ КРИВОЛИНЕЙНОЙ РЕГРЕССИИ

Даже в том случае, когда величины x и y не связаны между собой, наблюдаемые точки можно нанести на график и с помощью метода наименьших квадратов построить кривую, подобную (16.1.1) или (16.1.4). Найденные коэффициенты $\{b_i\}$ обычно не равны нулю. Требуется определить, когда коэффициенты не равны нулю, как в только что описанном случае, а когда это — следствие существующей зависимости между x и y.

Обычный подход заключается в построении таблицы дисперсионного анализа, применении F-критерия и проверке того, являются ли оцениваемые β-коэффициенты нулевыми. Если вычисленное F-отношение значимо велико, то мы отклоняем нулевую гипотезу, утверждающую, что все коэффициенты равны нулю, и заключаем, что регрессия значима.

В большинство регрессионных моделей входит коэффициент β_0 (среднее), отражающий функцию, не зависящую от x (например, модели (16.3.1) и (16.3.2)). Ненулевое значение этого коэффициента не предполагает наличие связи между x и y, поэтому мы не будем включать его в критерий значимости. Для проверки значимости других коэффициентов регрессии воспользуемся таблицей дисперсионного анализа 16.4.1.

Таблица 16.4.1. **Дисперсионный анализ. Криволинейная регрессия**

Источник вариации (1)	с. к. (2)	с. с. (3)	ср. к. (4) = (2)/(3)
Регрессия	$\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$\nu - 1$	(определяется делением)
Остаток	(определяется вычитанием)	$n - \nu$	
Общая вариация	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n - 1$	—

В этой таблице ν обозначает число β-коэффициентов. Проверяем, являются ли $\nu - 1$ из них (за исключением β_0) нулями. Критериальная статистика есть

$$F_{\nu-1,\,n-\nu} = \frac{\text{средний квадрат регрессии}}{\text{средний квадрат отклонений}}. \qquad (16.4.1)$$

Иногда необходимо проверить нулевую гипотезу о том, что все коэффициенты регрессии равны нулю (модель может не содержать параметр β_0 или необходимо проверить, являются ли β_0 и все остальные β-параметры нулями). В этом случае строится таблица нескорректированного[3] дисперсионного анализа (см. табл. 16.4.2), критериальная статистика в этом случае равна:

$$F_{\nu,\,n-\nu} = \frac{\text{средний квадрат регрессии}}{\text{средний квадрат отклонения}}, \qquad (16.4.2)$$

[3] Нескорректированность на среднее $\bar{y}$.

Таблица 16.4.2. **Нескорректированный дисперсионный анализ. Криволинейная регрессия**

Источник вариации (1)	с. к. (2)	с. с. (3)	ср. к. (4)=(2)/(3)
Регрессия	$\mathbf{b}'\mathbf{X}'\mathbf{y}$ (определяется вычитанием)	ν	(определяется делением)
Остаток		$n - \nu$	
Общая вариация	$\mathbf{y}'\mathbf{y}$	n	—

Пример 16.4.1. Проверим значимость коэффициентов регрессии в примере 16.1.1 с квадратичной функцией.

В данном примере $n = 4$, $\nu = 3$. Вычислим суммы квадратов, необходимые для дисперсионного анализа по формуле из табл. 16.4.1:

$$\text{общая с. к.} = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = \sum y_i^2 - n\bar{y}^2 = 101\,715 - 50\,850{,}25 = 50\,864{,}75,$$

$$\text{с. к. регрессии} = \mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2 = (89{,}79\,988;\ -113{,}12\,719;\ 30{,}636\,352) \times \begin{pmatrix} 451 \\ 2046 \\ 9534 \end{pmatrix} - 50\,850{,}25 = 101\,128{,}48 - 50\,850{,}25 = 50278{,}23,$$

$$\text{остаточная с. к.} = 50\,864{,}75 - 50\,278{,}23 = 586{,}52.$$

Читатель может проверить, что последнее значение есть сумма квадратов разностей между наблюдаемым и расчетным значениями y (см. параграф 15.5). Дисперсионный анализ приведен в табл. 16.4.3. Для проверки значимости регрессии вычисляем:

$$F_{2,\,1} = \frac{25\,139{,}11}{586{,}52} = 42{,}86.$$

Этот результат не значим при 5 %-ном уровне. При условии, что (16.3.1) — правильная модель, у нас нет доказательств того, что β_1 и β_2 не равны нулю.

Таблица 16.4.3. **Дисперсионный анализ данных из примера 16.4.1**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	50 278,23	2	25 139,11
Остаток	586,52	1	586,52
Общая вариация	50 864,75	3	—

Пример 16.4.2. Проверим на значимость коэффициент регрессии из примера 16.1.2 с экспоненциальной кривой.

В данном примере $n - 4$ и $\nu = 2$. Данные те же, что и в предыдущем примере, поэтому, как было подсчитано, общая сумма квадратов равна 50864,75. Для вычисления суммы квадратов регрессии воспользуемся формулами из таблицы дисперсионного анализа 16.4.1; остаточную сумму квадратов находим как разность. Итак,

$$\text{с. к. регрессии} = (-0{,}362\,598\,010;\ \mathbf{2{,}003631}\ 893)\begin{pmatrix}451\\ 50\,846{,}91\,948\end{pmatrix} -$$

$$- 50\,850{,}25 = 101\,714{,}9\,778 - 50\,850{,}25 = 50\,864{,}7\,278,$$

$$\text{остаточная с. к.} = 50\,864{,}75 - 50\,864{,}7\,278 = 0{,}0\,222.$$

Предоставляем читателю проверить, что последнее значение (с точностью до ошибок округления) равно сумме квадратов отклонений между наблюдаемыми и выравненными значениями y (см. параграф 15.5). Дисперсионный анализ представлен в табл. 16.4.4. Для проверки значимости регрессии вычислим

$$F_{1,\,2} = \frac{50\,864{,}7\,278}{0{,}0111} = 158 \times 10^4.$$

Этот результат весьма значим, поэтому имеется очевидное доказательство того, что коэффициент β_1 в модели (16.3.2) не равен нулю.

Таблица 16.4.4. **Дисперсионный анализ данных из примера 16.4.2**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	50 864,7278	1	50 864,7278
Остаток	0,0222	2	0,0111
Общая вариация	50 864,7500	3	—

Таблица 16.4.5. **Дисперсионный анализ данных из примера 16.4.3**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	101 714,71	1	101 714,71
Остаток	0,29	3	$9{,}66 \times 10^{-2}$
Общая вариация	101 715,00	4	—

Пример 16.4.3. Проверим значимость коэффициента регрессии из примера 16.2.3.

В данном случае $r = 4$ и $\nu = 1$, коэффициент β_0 отсутствует, поэтому необходимо воспользоваться формулами из табл. 16.4.2:

$$\text{общая с. к.} = \mathbf{y}'\mathbf{y} = \sum y_i^2 = 101\,715,$$

$$\text{с. к. регрессии} = \mathbf{b}'\mathbf{X}'\mathbf{y} = 2{,}000\,410\,580 \times 50\,846{,}91948 =$$

$$= 101\,714{,}71,$$

$$\text{остаточная с. к.} = 0{,}29.$$

Предоставляем читателю проверить, что последнее значение (с точностью до ошибок округления) равно сумме квадратов отклонений между наблюдаемыми и расчетными значениями y (см. параграф 15.5). Дисперсионный анализ приведен в параграфе 16.4.5. Для проверки значимости регрессии вычислим

$$F_{1,\,3} = \frac{101\,714{,}71}{9{,}66 \times 10^{-2}} = 1{,}05 \times 10^6.$$

Этот результат весьма значим. При условии, что модель выбрана правильно, можно утверждать, что имеется очевидное доказательство того, что β_1 не является нулем.

Литература: [19, с. 61—62; русский перевод с. 67—68].

16.5. ТОЧЕЧНЫЕ ОЦЕНКИ σ^2 И p

Параметры $\{\beta_i\}$ и σ^2 неизвестны, однако можно найти их несмещенные оценки. Коэффициенты регрессии $\{\beta_i\}$ оценивают $\{\beta_i\}$, а средний квадрат остатка в таблицах дисперсионного анализа 16.4.1 и 16.4.2 оценивают σ^2.

Ковариационная матрица $\{b_j\}$ равна $\sigma^2\mathbf{S}^{-1}$. Если элементы[4] обратной матрицы $\mathbf{S}^{-1}$ обозначить через $\{S_{ij}^{-1}\}$, то

$$\operatorname{var} v_i' = \sigma^2 S_{ii}^{-1}, \tag{16.5.1}$$

$$\operatorname{cov}(b_i,\ b_j) = \sigma^2 S_{ij}^{-1}. \tag{16.5.2}$$

Формулы (16.5.1) и (16.5.2) представляют собой обобщение формул (15.9.3), (15.9.4) и (15.9.5).

Литература: [19, с. 61; русский перевод с. 67—68].

16.6. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ σ^2 И β

Обозначим средний квадрат остатков в таблице дисперсионного анализа 16.4.1 или 16.4.2 через s^2, а элементы[5] обратной матрицы $\mathbf{S}^{-1}$ — через $\{S_{jj}^{-1}\}$. Для получения 95 %-ных доверительных границ для β_i приравняем (16.6.1) к верхней и нижней 2,5 %-ным точкам

[4] Если p-параметрами являются β_0, β_1, ..., то строки и столбцы S и S^{-1} имеют номера О, 1,...; если P-параметрами являются β_1, β_2,..., то строки и столбцы этих матриц имеют номера 1, 2,...

[5] См. сноску 4.

t_{n-v}-распределения (v здесь обозначает число оцениваемых параметров β):

$$t_{n-v} = (bi - \beta_i)/\left(s^2 S_{ii}^{-1}\right)^{\frac{1}{2}}. \qquad (16.6.1)$$

Для получения 95 %-ных доверительных границ для σ^2 приравняем

$$\chi^2_{n-v} = \frac{(n-v)\,s^2}{\sigma^2} \qquad (16.6.2)$$

к верхней и нижней 2,5 %-ной точке распределения χ^2_{n-v}. Формула (16.6.1) является обобщением формулы (15.9.6), а формула (16.6.2) — обобщением (15.7.3).

Пример 16.6.1. Найдем 95 %-ные доверительные интервалы для β_1 и σ^2 из примера 16.1.2.

В этом случае $n = 4$ и $v = 2$, средний квадрат остатка s^2 равен 0,0111 (см. табл. 16.4.4) и

$$\mathbf{S}^{-1} = \begin{pmatrix} 0{,}50154 & -0{,}44\,557 \times 10^{-2} \\ -0{,}44\,557 \times 10^{-2} & 0{,}78\,925 \times 10^{-4} \end{pmatrix}.$$

Для получения 95 %-ных доверительных границ для β_1 решим уравнения $(2{,}003\,631\,893 - \beta_1)/(0{,}0111 \times 0{,}789\,25 \times 10^{-4})^{1/2} = \pm 4{,}303$, откуда получим $\beta_1 = 2{,}003\,63 \pm 0{,}004\,03$.

Для получения 95 %-ных доверительных границ для σ^2 решим уравнения

$$\frac{2 \times 0{,}0111}{\sigma^2} = 7.38 \text{ и } \frac{2 \times 0{,}0111}{\sigma^2} = 0{,}0\,506,$$

откуда найдем, что границы равны 0,003 01 и 0,439.

Литература: [19, с. 64—65; русский перевод с. 74—75].

16.7. ПРОВЕРКА ГИПОТЕЗ

Для проверки гипотез о параметрах регрессии можно использовать t-статистику (16.6.1). Допустим, что необходимо проверить гипотезу $\beta_i = \delta$, где δ — некоторая заданная константа, причем модель содержит всего v параметров. Приравняем β_i к б в (16.6.1). Мы отклоняем нулевую гипотезу, если t-статистика попадает в верхнюю или нижнюю 2,5 %-ную область t_{n-v}-распределения (см. пример 15,8.1).

Очень часто нам требуется проверить нулевую гипотезу $\beta_i = 0$; t-статистику (16.6.1) можно применить и в этом случае, однако возможно использование и эквивалентного F-критерия, который может оказаться более удобным. Для этого необходимо проделать следующие вычисления:

1) подобрать модель с v параметрами и провести соответствующий дисперсионный анализ;

2) подобрать модель с $v - 1$ параметрами, содержащую все параметры, за исключением рассматриваемого b_i, и провести дисперсионный анализ. Сумма квадратов остатка в этом случае будет превышать остаточную сумму квадратов модели с v параметрами, а разность этих сумм связана с включением параметра b_i в уравнение;

3) найти значение критериальной статистики:

$$F_{1,\,n-v} = \frac{\text{уменьшение с. к. за счет } b_i}{\text{ср. к. для v параметров}}. \qquad (16.7.1)$$

Нулевая гипотеза отклоняется, если критериальная статистика попадает в верхнюю критическую область $F_{1,\,n-v}$-распределения.

Эта процедура проверки гипотез часто применяется в программах для ЭВМ по отысканию *уравнения «наилучшей» регрессии* (см. параграф 17.3). Необходимо заметить, что уменьшение суммы квадратов, связанное с включением члена с коэффициентом b_i, обычно зависит от того, какие переменные уже присутствуют в подгоняемом уравнении. Только в том случае, когда эти планы * взаимно ортогональны, уменьшение суммы квадратов остатка не будет зависеть от членов, включенных в регрессию ранее (см. параграф 16.13).

Вычислительная процедура, приводящая к (16.7.1),— частный случай более общего анализа эффекта введения q новых членов в модель, уже содержащую p членов. В этой ситуации дисперсионный анализ проводится в таблицах для p и $(p+q)$ параметров, результаты затем сводят в одну таблицу, как это показано в табл. 16.7.1.

Таблица 16.7.1. Дисперсионный анализ. Эффект введения q дополнительных членов в регрессионную модель с p членами

Источник вариации O	с. к. (2)	с. с. (3)	ср. к. (4) = (2)/(3)
Регрессии с p членами	(с. к. регрессии с p параметрами)	P	
Уменьшение, связанное с введением q новых членов	(определяется вычитанием)	q	(определяется делением)
Остаток	(остаточная с. к. модели с $(p+q)$ параметрами)	$n-p-q$	
Общая вариация	y'y	n	

Дополнительная сумма квадратов, связанная с включением q членов, равна разности между остаточными суммами квадратов в таблицах для p и $p+q$ параметров, а сумма квадратов остатка в формуле (16.7.1) есть остаточная сумма квадратов для модели с $p+q$ параметрами. Статистикой проверки того, приводят ли до-

* Имеются в виду $g_i(x)$ в уравнении (16.2.1).— *Примеч. пер.*

полнительные члены к значительно лучшему качеству подгонки, является

$$F_{q,\,n-p-q} = \frac{\text{изменение среднего квадрата}}{\text{средний квадрат отклонений}}. \qquad (16.7.2)$$

В табл. 16.7.1 представлен нескорректированный[6] дисперсионный анализ, основанный на табл. 16.4.2. Таблица скорректированного дисперсионного анализа построена так же, как табл. 16.4.2 (число степеней свободы в столбце (3) равно $p-1$, q, $n-p-q$ и $n-1$ соответственно).

Необходимо отметить, что критериальная статистика (16.7.1) может быть также использована для проверки нулевой гипотезы $\beta_i = 6$ (8 — обусловленная константа); при этом из каждого значения y_i надо вычесть δX_{ji} (X_{ji} есть j-й элемент i-го столбца матрицы X). Пример парной линейной регрессии был приведен в параграфе 15.8. Критерий проверки общей линейной гипотезы описан Дрейпером и Смитом в [19, с. 72—76; русский перевод с. 82—86].

Обращаем внимание читателя на параграф 12.2, где говорится об опасности применения нескольких критериев к одной совокупности данных.

Литература: [7, с. 441—446; русский перевод с. 317—320], [19, с. 67—68, 71—76, 163—194; русский перевод с. 76—78, 80—86, 159—164], [105, с. 269—271].

16.8. ОЖИДАЕМОЕ ЗНАЧЕНИЕ y ПРИ ЗАДАННОМ ЗНАЧЕНИИ x

Допустим, что, после того как $(k+1)$-параметрическая линия регрессии вида (16.2.1) построена, мы хотим оценить ожидаемое значение y, соответствующее заданному значению $X = x_0$. Для нахождения точечной оценки y_0 при условии, что $(k+1)$— вектор-строка

$$\mathbf{x}_0 = (g_0(x_0) g_1(x_0) \ldots g_k(x_0)) \qquad (16.8.1)$$

известна, воспользуемся формулой (15.10.2). (Размерность вектора x_0 равна размерности строки матрицы X). Для нахождения 95 %-ных доверительных границ среднего $\hat{y}$ приравняем

$$t_{n-k-1} = (\hat{y} - y_0) / \{s^2 (\mathbf{x}_0 \mathbf{S}^{-1} \mathbf{x}_0')\}^{\frac{1}{2}} \qquad (16.8.2)$$

к 2,5 %-ной точке t_{n-k-1}-распределения.

Обращаем внимание читателя на предостережение, сделанное в параграфе 15.10, об экстраполяции на основе регрессии.

Пример 16.8.1. Найдем 95 %-ные доверительные границы для среднего y в точке $x = 2$ для данных из примера 16.2.3 и табл. дисперсионного анализа 16.4.5. В этом случае $n = 4$, $k = 0$, векторы $\mathbf{b}$, $\mathbf{x}_0$ и S равны:

$$\mathbf{b} = 2{,}000\,410\,580, \quad \mathbf{x}_0 = e^2 = 7{,}38\,905,$$

$$\mathbf{S} = 25418{,}24\,162, \quad \mathbf{S}^{-1} = 1/25\,418{,}24\,162 = 3{,}93 \times 10^{-5}.$$

[6] См. параграф 16.4.

Средний квадрат остатка $s^2 = 0{,}0966$ (см. табл. 16.4.5). Из формулы (15.10.2) следует:

$$y_0 = \mathbf{x}_0 \mathbf{b} = 7{,}38905 \times 2{,}000410\,580 = 14{,}781.$$

Последнее значение и будет точечной оценкой среднего y при $x = 2$.

Для нахождения 95 %-ных доверительных границ для среднего y при $x = 2$ решим уравнения

$$(y - 14{,}781)/\{0{,}0\,966 \times (7{,}38\,905 \times 0{,}0\,000\,393 \times 7{,}38\,905)\}^{1/2} = \pm 3{,}182,$$

откуда получим $\hat{y} = 14{,}781 \pm 0{,}046$.

Литература: [19, с. 61; русский перевод с. 71].

16.9. ПРАВИЛЬНУЮ ЛИ МОДЕЛЬ МЫ ВЫБРАЛИ?

После того как линия регрессии построена, возникает вопрос: а правильную ли модель мы применили? Отсылаем читателя к параграфу 15.11. Критерий неадекватности подгонки, описанный в указанном параграфе, можно применять и для криволинейных регрессий (при этом, как и прежде, необходимо наличие нескольких наблюдений y для одного значения x). Вычислительная процедура[7] остается той же:

1) вычислить сумму квадратов регрессии (см. табл. 16.4.1);
2) вычислить общую сумму квадратов (см. табл. 10.4.1);
3) вычитанием найти остаточную сумму квадратов: (2) — (1);
4) вычислить сумму регрессии и определить неадекватность подгонки как сумму квадратов по формуле (15.11.2);
5) вычислить сумму квадратов неадекватности подгонки как разность: (4) — (1).
6) вычислить сумму квадратов чистой ошибки как разность: (3)-(5).

Соответствующие степени свободы показаны в табл. 16.9.1. В этой таблице v обозначает число коэффициентов регрессии, а p — число различных значений x. Неадекватность подгонки измеряется статистикой:

$$F_{p-v,\,n-p} = \frac{\text{средний квадрат неадекватности}}{\text{средний квадрат чистой ошибки}}. \qquad (16.9.1)$$

[7] В редких случаях требуется провести нескорректированный анализ по табл. 16.4.2. Тогда формулы в табл. 16.4.2. заменяются формулами из табл. 16.4.1 в расчетах неадекватности подгонки; член $T^2../n$ опускается из формулы (15.11.2). Степень свободы в табл. 16.9.1 остается без изменения, за исключением регрессии и общей с. к., которые теперь будут равны v и n соответственно (см. также пример 16.11.1).

Таблица **16.9.1. Дисперсионный анализ неадекватности подгонки. Криволинейная регрессия. Средний квадрат (обозначенный звездочками) получается делением суммы квадратов в колонке (2) на соответствующую степень свободы из колонки (3)**

Источник вариации (1)		с. к. (2)	с. с. (3)	ср. к. (4)=(2)/(3)
Регрессия		$\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$\nu - 1$	**
Остаток	Неадекватность подгонки	находится как разность $\sum_{i=1}^{p}(T_i^2/n_i) - T^2_{\cdot\cdot}/n$	$p - \nu$	**
	Чистая ошибка	находится как разность — с. к. регрессии	$n - p$	**
Общая вариация		$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n - 1$	—

Пример **16.9.1.** Подберем квадратичную кривую к данным из табл. 15.11.2 и проверим неадекватность подгонки.

Вид матрицы S и вектора X'y показан в (16.1.3). Суммируя различные степени $\{x_i\}$ и $\{y_i\}$, получим нормальные уравнения:

$$24b_0 + 805b_1 + 34\,225b_2 = 342,$$
$$805b_0 + 34\,225b_1 + 1\,651\,375b_2 = 13\,000,$$
$$34\,225b_0 + 1\,651\,375b_1 + 85\,793\,125b_2 = 574\,550,$$

где $b_0 = -2{,}262\,958\,508,$

$b_1 = 0{,}931\,520\,296,$

$b_2 = -0{,}010\,330\,543,$

а расчетные значения y при пяти различных значениях x равны 6,019; 12,235; 17,685; 18,469 и 16,438 соответственно. Подгонка оказалась хорошей. Вычислим теперь различные суммы квадратов:

с. к. регрессии $= \mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2 =$ (см. табл. 16.4.1),

$= 342b_0 + 13\,000b_1 + 574\,550b_2 - 4873{,}5 = 526{,}92,$

общая с. к. = 578,50 (см. пример 15.11.1),

остаточная с. к. = 578,50 — 526,92 = 51,98,

с. к. регрессии плюс с. к. неадекватности = 542,10 (см. пример 15.11.1),

с. к. неадекватности = 542,10 — 526,92 = 15,18,

с. к. чистой ошибки = 51,58 — 15,18 = 36,40.

Дисперсионный анализ приведен в табл. 16.9.2. Для проверки неадекватности регрессии вычислим

$$F_{2,19} = 7{,}6/1{,}9 = 4{,}0.$$

Эта величина значима при 5 %-ном уровне, но не значима при 2,5 %-ном уровне.

Таблица **16.9.2. Дисперсионный анализ. Пример 16.9.1**

Источник вариации		с. к.	с. с.	ср. к.
Регрессия		526,92	2	263,5
Остаток	Неадекватность	15,18	2	7,6
		51,58	21	2,5
	Чистая ошибка	36,40	19	1,9
Общая вариация		578,50	23	—

Некоторая неадекватность заметна и на рис. 16.9.1.

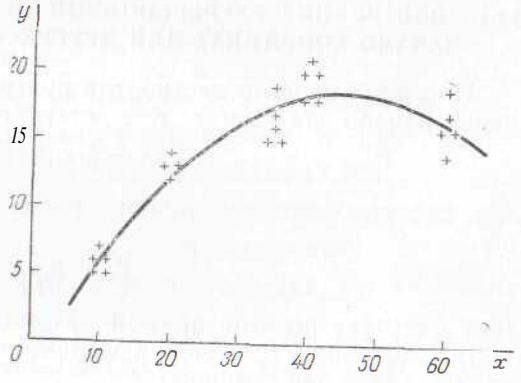


Рис. 16.9.1. Квадратичная кривая, построенная по методу наименьших квадратов

Литература: [19, с. 26—32, 86—100; русский перевод с. 34—40, с. 82—86].

16.10. НЕРАВНЫЕ ДИСПЕРСИИ. ВЗВЕШЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Эта проблема обсуждалась в параграфе 15.12 при рассмотрении парной линейной регрессии, поэтому мы отсылаем читателя к нему. Изложенные там методы могут быть применимы и к криволинейным моделям регрессии. Таким образом, если мы хотим определить только коэффициенты $\{b_i\}$, а дисперсия наблюдений в точке x_i равна σ^2/w_i, можно представить себе, что значение y_i наблюдалось Nw_i раз, где N достаточно большое целое число. Стандартный метод наименьших квадратов приведет нас тогда к нужным значениям $\{b_i\}$ (см. пример 15.12.1).

Данный метод *не может* применяться для статистического анализа. Для этого необходимо обобщить второй метод из параграфа 15.12 на криволинейные регрессии. В случае n наблюдений матрица X имеет n строк. Определим матрицу $\mathbf{X}_w$, которая получается из матрицы X домножением всех элементов первого ряда на $\sqrt{w_1}$, всех элементов второго ряда на $\sqrt{w_2}$ и т. д., и наконец, всех элементов n-го ряда на $\sqrt{w_n}$. Матрица $\mathbf{S}_w$, вектор $\mathbf{y}_w$ и среднее y_w определяются так же, как в параграфе 15.12. Затем, записывая $\mathbf{X}_w$ вместо X, $\mathbf{y}_w$ вместо y, $\mathbf{S}_w$ вместо S и подставляя вместо элементов $\mathbf{S}^{-1}$ элементы $\mathbf{S}_w^{-1}$ и y_w вместо y, можно применить стандартные формулы регрессионного анализа (15.3.1), (15.3.3), (15.3.4), (15.9.1), (15.10.2), табл. 16.4.2, (16.4.1), (16.4.2), (16.5.1), (16.5.2), (16.6.1), (16.6.2), табл. 16.7.1, (16.7.1), (16.7.2), (16.8.2). Если вместо n подставить $\sum w_i$, можно воспользоваться табл. 16.4.1, формулами (15.9.8), (15.9.9) и (15.9.10). Этот подход был описан в примере 15.12.2.

Литература: [19, с. 77—80; русский перевод с. 86—89].

16.11. ПОСТРОЕНИЕ РЕГРЕССИОННОЙ ПРЯМОЙ, ПРОХОДЯЩЕЙ ЧЕРЕЗ НАЧАЛО КООРДИНАТ ИЛИ ДРУГУЮ ФИКСИРОВАННУЮ ТОЧКУ

Иногда *априори* известно, что прямая регрессии должна пройти через начало координат, т. е. статистическая модель имеет вид:

$$y_i = \beta_1 x_i + e_i, \tag{16.11.1}$$

где $\{x_i\}$ свободны от ошибок. Тогда подгоняемая прямая имеет вид:

$$Y_i = b_1 x_i. \tag{16.11.2}$$

При подгонке по этой прямой матрица X будет иметь один столбец, содержащий n значений x. Векторы и матрицы регрессии тривиальны (все они скаляры):

$$\beta = \beta_1, \quad \mathrm{S} = \sum x_i, \quad \mathrm{X'y} = \mathrm{Z}\,\overline{x_i y_i},$$

$$b = b_1, \quad \mathrm{S}^{-1} = 1/(\sum x_i^2), \quad \mathrm{y'y} = \sum y_i^2.$$

В этом случае легко применить все матричные результаты из параграфов 16.2—16.10. Например, из параграфа 16.4 следует, что вектор b получается решением нормального уравнения Sb = X'y, поэтому

$$b_1 = (\sum x_i y_i)/(\sum x_i^2). \tag{16.11.3}$$

Из уравнения (16.5.1) следует:

$$\mathrm{var}\, b_1 = \sigma^2/(\sum x_i^2). \tag{16.11.4}$$

95 %-ный доверительный интервал для β_1 находят по формуле (16.6.1) с $\nu = 1$.

Читатель мог заметить, что дисперсионный анализ, соответствующий этому случаю, является нескорректированным и проводится, как в табл. 16.4.2 с $\nu = 1$. Коэффициент детерминации вычисляется делением суммы квадратов регрессии на общую сумму квадратов; полученное значение, вообще говоря, не будет квадратом обычного коэффициента корреляции между наблюдаемыми и расчетными значениями (как это было определено в (8.8.10)).

Иногда *априори* известно, что прямая регрессии должна пройти через заданную точку на плоскости $(\tilde{x}, \tilde{y})$. Наиболее простой способ построения прямой в этом случае заключается в том, чтобы из каждого значения x вычесть $\tilde{x}$, а из каждого значения y вычесть $\tilde{y}$ и проделать процедуру, описанную выше. Подогнанная прямая имеет вид:

$$(Y_i - \tilde{y}) = b_1 (x_i - \tilde{x}). \tag{16.11.5}$$

Пример 16.11.1. Подберем прямую линию вида (16.11.2), проходящую через 12 точек, взятых из табл. 16.11.1, и проверим ее на неадекватность. Начнем с вычисления

$$\sum x^2 = 255, \quad \sum xy = 266, \quad \mathrm{E}\, y^2 = 295.$$

Как следует из (16.11.3), $b_1 = 266/255 = 1{,}043\,137$. Нескорректированный дисперсионный анализ проводится, как показано в табл. 16.4.2. Итак, находим:

$$\text{с. к. регрессии} = \mathbf{b'X'y} = 1{,}043\,137 \times 266 = 277{,}475,$$

$$\text{общая с. к.} = \mathrm{y'y} = 295{,}000,$$

$$\text{остаточная с. к.} = 295{,}000 - 277{,}475 = 17{,}525.$$

В соответствии со сноской 7 на с. 293 и параграфом 16.9 сумма регрессии и сумма квадратов неадекватности равны:

$$2^2/3\ 4\ 4^2/3 + 14^2/3\ 4\ 25^2/3 = 280{,}333.$$

Таким образом,

$$\text{с. к. неадекватности} = 280{,}333 - 277{,}475 = 2{,}858,$$

$$\text{с. к. чистой ошибки} = 17{,}525 - 2{,}858 = 14{,}667.$$

Дисперсионный анализ показан в табл. 16.11.2.

Таблица 16.11.1. Данные для расчета регрессии для примера 16.11.1; 24 наблюдения взаимно независимы

x_i	y_{ij}			n_i	$T_{i\cdot}$
1	0	1	1	3	2
2	1	1	2	3	4
4	3	5	6	3	14
8	6	9	10	3	25
Всего				12	45

Таблица 16.11.2. Дисперсионный анализ данных из примера 16.11.1

Источник вариации		с. к.	с. с.	ср. к.
Регрессия		277,475	1	277,475
Остаток	Неадекватность	2,858	3	0,953
	Чистая ошибка	14,667	8	1,833
		17,525	11	1,593
Общая вариация		295,000	12	—

Для проверки на неадекватность вычислим

$$F_{3,8} = 0{,}953/1{,}833 = 0{,}520.$$

Эта величина незначима при 5 %-ном уровне, поэтому у нас нет доказательств неадекватности регрессии. Регрессия в высокой степени значима, поскольку

$$F_{1,11} = 277{,}475/1{,}593 = 174{,}2.$$

Для нахождения 95 %-ных доверительных границ для β_i решим уравнения

$$(1{,}043\ 137 - \beta_1)/\{1{,}593 \times (1/255)\}^{1/2} = \pm 2{,}201,$$

откуда найдем $\beta_1 - 1{,}043 \pm 0{,}174$.

Литература: [7, с. 358—361; русский перевод с. 333—336], [102, с. 273], [105, с. 214—215].

16.12. ОРТОГОНАЛЬНЫЙ ПОЛИНОМИАЛЬНЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Допустим, нам необходимо для совокупности точек (x_i, y_i), $i = 1, 2, \ldots$, подобрать полиномиальную кривую. Можно начать с прямой линии, применяя методы, рассмотренные в гл. 15. Если результаты подгонки будут неудовлетворительны, то можно попытаться, применяя метод из примера 16.1, перейти к квадратичной зависимости. Если и в этом случае кривая будет плохо подогнанной, то можно перейти к кубической зависимости и т. д. При этом каждый раз следует пересчитывать все коэффициенты регрессии. Объем связанной с этим работы весьма значителен.

Если $\{x_i\}$ распределены равномерно, нет необходимости каждый раз пересчитывать коэффициенты регрессии. Применяя *ортогональные полиномы (Чебышева)*, можно значительно сократить время и затрачиваемые усилия. Далее будем предполагать, что значения x стандартизованны, т. е. они распределены равномерно с единичным интервалом изменения и симметрично относительно начала координат. Например, при $n = 3$ значения x будут равны -1, 0 и 1, при $n = 4$ они равны $-\frac{3}{2}$, $-\frac{1}{2}$, $\frac{1}{2}$ и $\frac{3}{2}$, при $n = 5$ равны -2, -1, 0, 1 и 2.

Обозначим ортогональный полином степени r через $\varphi_r(x)$. Эти полиномы обладают одним важным свойством: если $\{x_i\}$ стандартизованны и $r \neq s$, то

$$\sum_{i=1}^{n} \varphi_r(x_i)\,\varphi_s(x_i) = 0. \qquad (16.12.1)$$

Это свойство единственным образом определяет все ортогональные полиномы, начиная с константы. Можно показать [8], что

$$\varphi_0(x) = 1, \qquad (16.12.2)$$

$$\varphi_1(x) = \lambda_1 x, \qquad (16.12.3)$$

$$\varphi_2(x) = \lambda_2 \left\{x^2 - \frac{1}{12}(n^2 - 1)\right\}, \qquad (16.12.4)$$

$$\varphi_3(x) = \lambda_3 \left\{x^3 - \frac{1}{20}(3n^2 - 7)\,x\right\}. \qquad (16.12.5)$$

Полиномы более высокого порядка приводятся, например, в [76, с. 212]*. Константы $\{\lambda_r\}$ обычно выбираются так, чтобы ортогональные полиномы в стандартизованных точках принимали целые значения. Значения ортогональных полиномов (в стандартизованных точках) для степени 4 и менее при $n = 13$ приведены в табл. 16.12.1. Эта таблица взята из [76, с. 212—221], читатель может проверить, что условие (16.12.1) здесь выполнено.

Любой полином k-й степени может быть выражен как линейная комбинация ортогональных полиномов степени k и меньше. Таким образом, для подгонки полинома k-й степени мы можем использовать кривую вида

$$Y = b_0\varphi_0(x) + b_1\varphi_1(x) + \ldots + b_k\varphi_k(x). \qquad (16.12.6)$$

Удобство перехода к (16.12.6) очевидно. Значение j-го столбца $(j = 0, 1, \ldots, k)$ матрицы X равно значению ортогонального полинома j-й степени в стандартизованных точках, а элементы матрицы $S = X'X$ даются выражением (16.2.3) при подстановке ф вместо g. Уравнение (16.12.1) означает, что S_{rs} и S_{sr} равны нулю при $r \neq s$, т. е. матрица S диагональная. Обратная матрица S^{-1} получается заменой каждого диагонального элемента матрицы S обратным ему (см. параграф 1.8). В случае $n = 13$ диагональные элементы матрицы S в табл. 16.12.1 представляют последнюю строку этой таблицы. Для нахождения j-го элемента $(j = 0, 1, \ldots, k)$ вектора $X'Y$ необходимо каждое значение y домножить на соответствующее значение ортогонального полинома j-й степени, а произведения сложить.

[8] См. упражнение 7 из параграфа 16.14.

* См. также: Суетин С. П. Классические ортогональные многочлены. М., Наука, 1976.— *Примеч. пер.*

Таблица 16.12.1. **Численные значения (в стандартизованных точках) ортогональных полиномов степени 4 и менее для** $n=13$

	$\varphi_0(x)$	$\varphi_1(x)$	φ_2 O)	$\Phi_3(x)$	$\Phi_4(x)$
	1	-6	22	-11	99
	1	-5	11	0	-66
	1	-4	2	6	-96
	1	-3	-5	8	-54
	1	-2	-10	7	11
	1	-1	-13	4	64
	1	0	-14	0	84
	1	1	-13	-4	64
	1	2	-10	-7	11
	1	3	-5	-8	-54
	1	4	2	-6	-96
	1	5	11	0	-66
	1	6	22	11	99
λ_r	1	1	1	1/6	7/12
$\sum_{i=1}^{n} \{\varphi_r(x_i)\}^2$	13	182	2002	572	68 068

Источник. [76]. Перепечатано с разрешения Biometrika Trustees.

Применение ортогональных полиномов в регрессионном анализе было развито Р. А. Фишером [23] в 1921 г. Позднее он предложил использовать коэффициенты пропорциональности $\{\lambda_i\}$ с тем, чтобы полиномы в стандартизованных точках принимали целые значения.

Пример 16.12.1. С помощью метода ортогональных полиномов подберем кривую нулевого, первого, второго и третьего порядков к 13 точкам, взятым из табл. 16.12.2.

Таблица 16.12.2. **Данные для ортогональных полиномов из примера 16.12.1. Расчетные значения $\{Y_i\}$ основаны на кубической параболе, найденной по методу наименьших квадратов**

x_i	y_i	Y_i	x_i	y_i	Y_i
10,1	30,3	30,30	10,8	259,7	259,69
10,2	61,2	61,20	10,9	295,0	295,00
10,3	92,7	92,72	11,0	331,0	330,98
10,4	124,9	124,86	11,1	367,6	367,61
10,5	157,6	157,61	11,2	404,9	404,91
10,6	191,0	191,00	11,3	442,9	442,89
10,7	225,02	225,02			

Если значения x-в стандартизованы, элементы вектора X'y можно найти следующим образом:

$$\sum y_i\varphi_0(x_i) = 30{,}3 + 61{,}2 + 92{,}7 + \ldots + 442{,}9 = 2983{,}8,$$
$$\sum y_i\varphi_1(x_i) = 30{,}3 \times (-6) + 61{,}2 \times (-5) + \ldots + 442{,}9 \times 6 = 6255{,}5,$$
$$\sum y_i\varphi_2(x_i) = 30{,}3 \times 22 + 61{,}2 \times 11 + \ldots + 442{,}9 \times 22 = 643{,}5,$$
$$\sum y_i\varphi_3(x_i) = 30{,}3 \times (-11) + 61{,}2 \times 0 + \ldots + 442{,}9 \times 11 = 3{,}8.$$

Диагональные элементы S^{-1} являются обратными к последней строке табл. 16.12.1, а именно они равны 1/13, 1/182, 1/2002, 1/572. Недиагональные элементы равны нулю. Поскольку $b = S^{-1}X'y$, то

$$b_0 = 2983{,}8/13 = 229{,}5\,230\,769,$$
$$b_1 = 6255{,}5/182 = 34{,}37\,087\,912,$$
$$b_2 = 643{,}5/2002 = 0{,}321\,428\,571,$$
$$b_3 = 3{,}8/572 - 0{,}006\,643\,356.$$

Вспомним теперь, что значения x не стандартизованны; среднее наименьшего и наибольшего значений равно 10,7, а промежуток между соседними значениями x равен 0,1. Стандартизованные значения x, таким образом, получаются вычитанием 10,7 и делением результата на 0,1. Принимая это во внимание, а также с учетом формул (16.12.2) — (16.12.5) и значений λ из табл. 16.12.1, определим окончательно вид найденных по методу наименьших квадратов аппроксимирующих полиномоь нулевой, первой, второй и третьей степеней.

Полином нулевого порядка, найденный по методу наименьших квадратов, имеет вид:

$$Y = 229{,}5\,230\,769,$$

прямая метода наименьших квадратов имеет вид:

$$У = 229{,}5\,230\,769 + 34{,}37\,087\,912\left(\frac{x - 10{,}7}{0{,}1}\right).$$

Квадратичная функция метода наименьших квадратов равна:

$$Y = 229{,}5\,230\,769 + 34{,}37\,087\,912\left(\frac{x - 10{,}7}{0{,}1}\right) + 0{,}321\,428\,571\left\{\left(\frac{x - 10{,}7}{0{,}1}\right)^2 - 14\right\},$$

кубическая парабола имеет вид:

$$Y = 229{,}5\,230\,769 + 34{,}37\,087\,912\left(\frac{x - 10{,}7}{0{,}1}\right) + 0{,}321\,428\,571\left\{\left(\frac{x - 10{,}7}{0{,}1}\right)^2 - 14\right\} + \frac{1}{6}(0{,}006\,643\,356) \times \left\{\left(\frac{x - 10{,}7}{0{,}1}\right)^3 - 25\left(\frac{x - 10{,}7}{0{,}1}\right)\right\}.$$

Вычисление расчетных значений в стандартизованных точках можно проводить непосредственно с помощью табл. 16.12.1. Например, выравненное значение y в четвертой точке, вычисленное на основе кубической параболы, будет равно:

$$229{,}5\ 230\ 769 + 34{,}37\ 087\ 912 \times (-3) + 0{,}321\ 428\ 571 \times (-5) + \\ + 0{,}006\ 643\ 356 \times 8 = 124{,}86.$$

Литература: [9, с. 173], [19, с. 69—70, с. 150—155; русский перевод с. 78—79, с. 159—164], [54, с. 359—362], [66, с. 131—134], [76, с. 91—95, с. 212—221].

16.13. ОРТОГОНАЛЬНАЯ ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ. СТАТИСТИЧЕСКИЙ АНАЛИЗ

Как было показано в предыдущем параграфе, применение ортогональных полиномов приводит к тому, что матрица S становится диагональной. Обратная матрица $\mathbf{S}^{-1}$ в этом случае может быть найдена очень просто, что значительно сокращает время вычислений и экономит усилия при отыскании полиномов по методу наименьших квадратов. Это же удобство сохраняется и для статистического анализа модели ортогональных полиномов:

$$y_i = \beta_0\varphi_0(x_i) + \ldots + \beta_k\varphi_k(x_i) + e_i. \qquad (16.13.1)$$

Значения $\{x_i\}$ не содержат ошибки, а $\{e_i\}$ являются независимыми нормально распределенными случайными величинами с нулевым математическим ожиданием и одинаковой дисперсией σ^2.

Как и прежде, можно построить необходимые векторы и матрицы. Многие из них теперь существенно упрощаются. Так, $\mathbf{S}^{-1}$ будет диагональной матрицей, а в соответствии с формулой (16.5.2) это влечет некоррелируемость коэффициентов регрессии. Коэффициенты регрессии будут нормально распределены, поэтому можно утверждать, что они стохастически независимы (см. параграф 9.6).

Дисперсионный анализ проводится обычно для нескорректированного случая (см. параграф 16.4). Каждый дополнительный полиномиальный член приводит к лучшей подгонке и уменьшению суммы квадратов отклонений. Поскольку полиномы ортогональны, уменьшение суммы квадратов, связанное с введением нового полинома, не зависит от полиномов, участвовавших в подгонке ранее. Сумма квадратов регрессии $\mathbf{b}'\mathbf{X}'\mathbf{y}$, таким образом, может быть разбита единственным образом на независимые слагаемые, отвечающие разным полиномам. Эти слагаемые вносят индивидуальный эффект в сумму $\mathbf{b}'\mathbf{X}'\mathbf{y}$. Уменьшение суммы квадратов, соответствующее, например, полиному j-го порядка, равно

$$b_j\left(\sum_{i=1}^{n} y_i\varphi_j(x_i)\right).$$

Дисперсионный анализ регрессии представлен в табл. 16.13.1.

Таблица 16.13.1. Дисперсионный анализ. Ортогональная полиномиальная регрессия

Источник вариации (1)	с. к. (2)	с. с, (3)	ср. к. (4)=(2)/(3)
Среднее (b_0)	$b_0\sum_{i=1}^{n} y_i$	1	Г то же, что в] [столбце (2) \
Линейный член (b_1)	$b_1\sum_{i=1}^{n}$ Уда (x_i)	1	
Квадратичный член (b_2)	$b_2\sum_{i=1}^{n} y_i\varphi_2(x_i)$	1	
. . .	. . .	. . .	
Член порядка k (b_k)	$b_k\sum_{i=1}^{и} y_i\varphi_k(x_i)$	1	
Остаток	(определяется вычитанием)	$n-k-1$	(определяется делением)
Общая вариация	$\sum_{i=1}^{n} y_i^2$	n	—

Значимость коэффициента регрессии b_j проверяется по F-статистике:

$$F_{1,\,n-k-1} = \frac{\text{средний квадрат}}{\text{средний квадрат отклонения}}. \qquad (16.13.2)$$

Критерии значимости, подобные (16.13.2), применяются для нахождения степени полинома, по которому производится подгонка, а также для определения, какой из членов более низкого порядка может быть опущен. Часто применяют последовательные процедуры. Отметим, однако, что даже если коэффициент β_j равен нулю или очень мал, то коэффициент β_{j+k} $(k = 1, 2, \ldots)$ при полиноме более высокого порядка может быть большим. В этом случае остаточная сумма квадратов, найденная для полинома j-й степени, не будет соответствовать верному значению среднего квадрата отклонений для F-критерия при проверке b_j. В связи с этим здесь необходима осторожность. Если мы готовы предположить, что подгоняемая функция имеет степень не выше k, то дисперсионный анализ в табл. 16.13.1 и вычисления F-статистики (16.13.2) для проверки значимости b_j должны быть произведены для всех $j \leqslant k$. Незначимые члены исключаются из регрессионного уравнения, а соответствующие суммы квадратов добавляются к остаточной сумме квадратов для последующего анализа (например, для оценивания). Число степеней свободы для остатков при этом соответственно увеличивается.

Пример 16.13.1. Применим метод ортогональных полиномов для подгонки полинома низкого порядка к 13 точкам, взятым из табл. 16.12.2.

Допустим, подгоняемый полином имеет порядок 4 или менее. Из примера 16.12.1, известно, что

$$\begin{aligned} \sum y_i\varphi_0(x_i) &= 2983{,}8, & b_0 &= 229{,}523\ 07\ 69,\\ \sum y_i\varphi_1(x_i) &= 6255{,}5, & b_1 &= 34{,}370\ 879\ 12,\\ \sum y_i\varphi_2(x_i) &= 643{,}5, & b_2 &= 0{,}321\ 428\ 571.\\ \sum y_i\varphi_3(x_i) &= 3{,}8, & b_3 &= 0{,}006\ 643\ 356,\end{aligned}$$

Легко подсчитать, что

$$\sum y_i\varphi_4(x_i) = 0{,}2, \qquad b_4 = 0{,}000\ 002\ 938,$$

$$\sum y_i^2 = 900\,064{,}86.$$

Уменьшения в остаточной сумме квадратов полиномов нулевой, первый и т. д. степеней соответственно равны:

$$\begin{aligned} b_0&: 229{,}523\ 076\ 9 \times 2\ 983{,}8 = 684\ 850{,}9568,\\ b_1&: 34{,}370\ 879\ 12 \times 6\ 255{,}5 = 215\ 007{,}0343,\\ b_2&: 0{,}321\ 428\ 571 \times 643{,}5 = 206{,}8393,\\ b_3&: 0{,}006\ 643\ 356 \times 3{,}8 = 0{,}0252,\\ b_4&: 0{,}000\ 002\ 938 \times 0{,}2 = 5{,}88 \times 10^{-7}.\end{aligned}$$

Дисперсионный анализ и значения F-статистик показаны в табл. 16.13.2. Коэффициент b_4 незначим, а все остальные (в том числе b_3) значимы с высокой степенью. Поэтому подгоняемая функция — кубическая парабола из примера 16.12.1 — содержит $\varphi_0(x)$, $\varphi_1(x)$, $\varphi_2(x)$ и $\varphi_3(x)$. Расчетные значения показаны в табл. 16.12.2.

Таблица **16.13.2. Дисперсионный анализ. Пример 16.13.1 с ортогональными полиномами**

Источник вариации	с. к.	с. с.	ср. к.	p
Среднее (b_0)	684 850,9568	1	$6{,}85 \times 10^5$	$1{,}25 \times 10^9$
Линейный член (b_1)	215 007,0343	1	$2{,}15 \times 10^5$	$3{,}91 \times 10^3$
Квадратичный член (b_2)	206,8393	1	$2{,}07 \times 10^2$	$3{,}76 \times 10^0$
Кубический член (b_3)	0,0252	1	$2{,}52 \times 10^{-2}$	$4{,}58 \times 10^1$
Член 4-й степени (64)	0,0000 *	1	$5{,}88 \times 10^{-7}$	$1{,}07 \times 10^{-3}$
Остаток	0,0044	8	$5{,}50 \times 10^{-4}$	—
Общая вариация	900 064,8600	13	—	—

* $5{,}88 \times 10^{-7}$.

Доверительные интервалы для β_0, β_1, β_2, РЗ и σ^2 могут быть найдены с помощью статистик (16.6.1) и (16.6.2). Заметим, что

$$\delta_{00}^{-1} = \frac{1}{13}, \quad \delta_{22}^{-1} = \frac{1}{2002},$$

$$S_{11}^{-1} = \frac{1}{182}, \quad S_{33}^{-1} = \frac{1}{572}.$$

Коэффициент при члене 4-й степени незначим и остаточный средний квадрат s^2 может быть получен сложением суммы квадратов, связанной с включением члена четвертого порядка, с полученным ранее остатком и делением результата на 9 степеней свободы вместо 8. Теперь обратимся к статистикам (16.6.1) и (16.6.2) с $n = 13$, $\nu - 4$.

Литература: [19, с. 150—155; русский перевод с. 159—164], [66, с. 131 — 134], [76, с. 91—95, 212—221].

16.14. УПРАЖНЕНИЯ

1. Способом, описанным в параграфе 15.2, получите нормальные уравнения для подгонки полинома второй степени (16.1.1) по методу наименьших квадратов к четырем точкам (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4). Примите во внимание формулы (16.1.3).

2. Способом, описанным в параграфе 15.2, получите нормальные уравнения для подгонки экспоненциальной кривой (16.1.4) по методу наименьших квадратов к четырем точкам (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4). Примите во внимание формулы (16.1.6).

3. Какая из приведенных моделей является линейной для метода наименьших квадратов (см. параграф 16.2)?

а) $y_i = b_0 + b_1\sqrt{x_i} + e_i$;

б) $y_i = b_1(1 + x_i)^{\frac{3}{2}} + e_i$;

в) $y_i = b_1 \sin(b_2 x_i)$;

г) $y_i = b_0 + b_1/x_i + b_2/x_i^2 + e_i$.

4. Преобразуйте следующие статистические модели так, чтобы было возможным применение стандартного линейного метода наименьших квадратов:

а) $y_i = a_1(x_i - a_0) + e_i$;

б) $y_i = k \sin(\pi x_i + o.) + e_i$.

5. С помощью метода наименьших квадратов при матричном обозначении подберите полином третьей степени к четырем точкам из примера 16.1.1.

6. Подберите кривую вида $Y = b_0 + b_1 x^3$ к данным из -табл. 16.12.2. Проверьте регрессионный коэффициент b_1 на значимость.

7. Ортогональный полином нулевого порядка для пяти стандартизованных точек -2, -1, О, 1, 2 равен $\varphi_0(x) = 1$. Предположим, что $\varphi_1(x) = Ax + B$. При условии ортогональности (16.12.1) имеем:

$$\sum_{x=-2}^{2} \varphi_0(x)\,\varphi_1(x) = \sum_{x=-2}^{2} (Ax + B) = 5B = 0.$$

Тогда B равно нулю и $\varphi_1(x)$ пропорционален x, таким образом, можно записать $\varphi_1(x)=\lambda_1 x$. Предположим, $\varphi_2(x)=ax^2+bx+c$, и, учитывая, что $\varphi_2(x)$ должен быть ортогонален одновременно и к $\varphi_0(x)$, и к $\varphi_1(x)$, можно установить вид $\varphi_2(x)$ при $n=5$. Примените этот же метод для нахождения $\varphi_3(x)$.

17. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

В этой главе будет показано, как матричные методы, рассмотренные в гл. 15 и 16, можно применять при изучении регрессии зависимой переменной y на две или более независимые переменные. Коротко обсудим также вопрос о выборе «наилучшей» регрессии.

17.1. ВВЕДЕНИЕ. НЕСКОЛЬКО ПРОСТЫХ ПРИМЕРОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

В гл. 15 и 16 были описаны методы парной линейной регрессии и криволинейной регрессии зависимой переменной на одну независимую переменную x. Та же техника может быть использована при изучении регрессии зависимой переменной на две или более независимые переменные x, z и т. д., причем предполагается, что уравнение регрессии линейно[1] по коэффициентам $\{b_i\}$. Такая регрессия называется *множественной линейной регрессией*. Матричные методы здесь становятся необходимыми; все формулы и методы были даны ранее, ссылки на них для удобства собраны в табл. 17.1.1.

Таблица 17.1.1. **Множественная линейная регрессия (ссылки)**

Техника, метод	Ссылка
1. Метод наименьших квадратов:	
минимизация функции	(15.3.1)
формула для нахождения выравненных значений	(15.3.2)
нормальные уравнения	(15.3.3)
2. Статистическая модель	(15.9.1)
3. Проверка значимости регрессии	параграф 16.4
Скорректированный дисперсионный анализ	табл. 16.4.1, (16.4.1)
Нескорректированный дисперсионный анализ	табл. 16.4.2, (16.4.2)
4. Точечные оценки для $\{\beta_i\}$ и σ^2	параграф 16.5
Ковариационная матрица $\{b_i\}$	(16.5.1), (16.5.2)
5. Доверительные интервалы для $\{\beta_i\}$ и σ^2	параграф 16.6
6. Проверки гипотез	параграф 16.7, пример 17.1.3
7. Ожидаемое значение y при данном значении независимой переменной — точечная оценка y_0	(15.10.2), пример 17.1.2
Доверительные границы для $\hat{y}$	(16.8.2), пример 17.1.2
8. Неадекватность подгонки	параграф 16.9, пример 17.1.5
9. Взвешенный метод наименьших квадратов	параграф 16.10, пример 17.1.4

[1] См. параграф 16.2.

В примерах 17.1.1—17.1.5 показывается, как строятся матрицы и векторы регрессии y, b, X и S. В этих примерах также демонстрируется применение следующих методов регрессионного анализа: скорректированного и нескорректированного дисперсионного анализа при проверке значимости регрессии (примеры 17.1.1 и 17.1.3), доверительных границ для параметров регрессии $\{\beta_i\}$ и σ^2 (пример 17.1.1), проверки гипотез (пример 17.1.3), доверительных границ для ожидаемого значения зависимой переменной при заданном значении независимой переменной (пример 17.1.2), проверки на неадекватность (пример 17.1.5), взвешенного метода наименьших квадратов (пример 17.1.4).

Пример 17.1.1. С помощью метода наименьших квадратов подберем кривую вида

$$Y=b_0+b_1x+b_2e^z \qquad (17.1.1)$$

к четырем (трехмерным) точкам:

$$(x_1,\ z_1,\ y_1)=(3;\ 1;\ 8{,}2),$$

$$(x_2,\ z_2,\ y_2)=(20;\ 3;\ 60{,}3),$$

$$(x_3,\ z_3,\ y_3)=(1;\ 0;\ 3{,}1),$$

$$(x_4,\ z_4,\ y_4)=(55;\ 4;\ 164{,}3).$$

Независимыми переменными в данной регрессии являются x и z, зависимой переменной — y. Проверим регрессионное уравнение на значимость и найдем доверительные границы для неизвестного параметра Ро и σ^2.

Прямая линия содержит нулевую и первую степень x, а при построении прямой линии по методу наименьших квадратов (см. параграф 15.3) мы используем матрицу X с двумя столбцами, один из которых составлен из нулевых степеней $\{x_i\}$, а другой — из первых степеней $\{x_i\}$. При подгонке по экспоненциальной кривой (16.1.4), содержащей константу и член, пропорциональный ряду e^x, возьмем матрицу X также с двумя столбцами, один из которых составлен из единиц, а второй — из значений $\{\exp x_i\}$. Для подгонки по кривой вида (17.1.1) матрица X должна иметь уже три столбца. Первый, второй и третий столбцы составлены соответственно из единиц, значений $\{x_i\}$ и значений $\{\exp z_i\}$. Как обычно, $\{y_i\}$ образуют вектор-столбец y, а коэффициенты регрессии — трехмерный вектор-столбец b. Итак,

$$\mathrm{X}=\begin{pmatrix}1 & 3 & 2{,}718\,281\,828\\ 1 & 20 & 20{,}085\,536\,692\\ 1 & 1 & 1{,}000\,000\,000\\ 1 & 55 & 54{,}598\,150\,03\end{pmatrix} \quad \text{и} \quad \mathrm{y}=\begin{pmatrix}8{,}2\\ 60{,}3\\ 3{,}1\\ 164{,}3\end{pmatrix}.$$

Поэтому

$$\mathbf{S} = \mathbf{X'X} = \begin{pmatrix} 4 & 79 & 78{,}401\,968\,78 \\ 79 & 3435 & 3413{,}763\,836 \\ 78{,}401\,968\,79 & 3413{,}763836 & 3392{,}775\,836 \end{pmatrix}$$

и

$$\mathbf{X'y} = \begin{pmatrix} 235{,}9 \\ 10\,270{,}2 \\ 10\,207{,}023\,84 \end{pmatrix}.$$

Нормальные уравнения (15.3.3) имеют вид:

$$4b_0 + 79b_1 + 78{,}401\,968\,78b_2 = 235{,}9,$$

$$79b_0 + 3435b_1 + 3413{,}763\,836b_2 = 10\,270{,}2,$$

$$78{,}401\,968\,78b_0 + 3413{,}763\,836b_1 + 3392{,}775\,836b_2 = 10\,207{,}023\,84.$$

Находим: $b_0 = -0{,}001\,022\,462$, $b_1 = 0{,}293\,743\,778$ и $b_2 = 2{,}712\,920\,796$, поэтому выравненные значения равны соответственно 8,255; 60,364; 3,006 и 164,275.

Статистическая модель (15.9.1) в этом случае имеет вид:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \exp z_i + e_i, \qquad (17.1.2)$$

где $\{x_i\}$ и $\{z_i\}$ не содержат ошибок. Случайные величины $\{e_i\}$ независимы, имеют нормальное распределение с математическим ожиданием, равным нулю, и одинаковой дисперсией σ^2. Для проверки значимости коэффициентов b_1 и b_2 воспользуемся дисперсионным анализом (см. табл. 16.4.1) при $\nu = 3$. Тогда

$$\text{общая с. к.} = \mathbf{y'y} - n\bar{y}^2 = \sum y_i^2 - 4(235{,}9/4)^2 =$$

$$= 30\,707{,}43 - 13\,912{,}2025 = 16\,795{,}2275,$$

$$\text{с. к. регрессии} = \mathbf{b'X'y} - n\bar{y}^2 = (b_0 b_1 b_2)\begin{pmatrix} 235{,}9 \\ 10270{,}2 \\ 10\,207{,}023\,84 \end{pmatrix} -$$

$$- 13\,912{,}2025 = 16795{,}2109.$$

Дисперсионный анализ[2] приведен в табл. 17.1.2. Критериальная статистика

$$F_{2,1} = 8397{,}6054/0{,}0166 = 5{,}06 \times 10^5$$

значима в высокой степени. Если модель (17.1.2) выбрана правильно, то можно утверждать, что имеется достаточно оснований считать b_1 и b_2 не равными нулю.

[2] С точностью до ошибок округления остаточная сумма квадратов равна сумме квадратов разностей между наблюдаемыми и расчетными значениями.

Таблица 17.1.2. Дисперсионный анализ. Пример множественной регрессии **17.1.1**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	16 795,2109	2	8397,6054
Остаток	0,0166	1	0,0166
Общая вариация	16 795,2275	3	—

Доверительный интервал для β_0 может быть определен с помощью (16.6.1) при $n = 4$ и $\nu = 3$. Находим

$$\mathbf{S}^{-1} = \begin{pmatrix} 0{,}480\,70 & -0{,}459\,17 & 0{,}450\,90 \\ -0{,}459\,17 & 8{,}88950 & -8{,}933\,88 \\ 0{,}450\,90 & -8{,}93388 & 8{,}979\,02 \end{pmatrix},$$

средний квадрат отклонений $s^2 = 0{,}0166$. Решим теперь уравнения

$$(-0{,}001\,022\,462 - \beta_0)/(0{,}0166 \times 0{,}480\,70)^{1/2} = \pm 12{,}706,$$

откуда найдем 95 %-ные доверительные границы для β_0: $-0{,}001\,02 \pm 1{,}135$.

Доверительные границы для σ^2 получают по формуле (16.6.2). Решим следующие уравнения:

$$\frac{(4-3)\,0{,}0166}{\sigma^2} = 5{,}02 \text{ и } \frac{(4-3)\,0{,}0166}{\sigma^2} = 0{,}000\,982,$$

откуда 95 %-ные доверительные границы равны 0,00331 и 16,9. Доверительные интервалы широки из-за малого объема выборки.

Пример 17.1.2. С помощью метода наименьших квадратов подберем кривую вида

$$Y = b_0 + b_1 (x e^z)^{\frac{1}{2}} \qquad (17.1.3)$$

к четырем точкам из примера 17.1.1. Проверим значимость регрессии и найдем 95 %-ный доверительный интервал для ожидаемого значения y при $x = 10$ и $z = 2$.

В данном случае матрица X будет иметь два столбца: первый составлен из единиц, а второй — из значений $\sqrt{x_i \exp z_i}$. Матрица $\mathbf{S} = \mathbf{X'X}$, таким образом, равна:

$$\begin{pmatrix} n & \sum (x_i \exp z_i)^{\frac{1}{2}} \\ \sum (x_i \exp z_i)^{\frac{1}{2}} & \sum x_i \exp z_i \end{pmatrix} =$$

$$= \begin{pmatrix} 4 & 78{,}697\,098\,47 \\ 78{,}697\,098\,47 & 3413{,}763\,833 \end{pmatrix}$$

и

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum y_i (x_i \exp z_i)^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} 235{,}9 \\ 10\,238{,}520\,17 \end{pmatrix}.$$

Нормальные уравнения (15.3.3) имеют следующий вид:

$$4b_0 + 78{,}697\,098\,47 b_1 = 235{,}9,$$

$$78{,}697\,098\,47 b_0 + 3413{,}763\,833 b_1 = 10\,238{,}520\,17,$$

откуда $bo = -0{,}058\,288\,883$ и $b_1 = 3{,}000531\,915$. Выравненные значения y равны 8,510; 60,081; 2,942; 164,367.

Для проверки значимости коэффициента регрессии b_1 воспользуемся таблицей дисперсионного анализа[3] 16.4.1 с $n = 4$ и $\nu = 2$. Общая сумма квадратов уже известна и равна 16795, 2275, а

$$\text{с. к. регрессии} = \mathbf{b}'\mathbf{X}'\mathbf{y} - ny^2 = (b_0 \; b_1) \begin{pmatrix} 235{,}9 \\ 10\,238{,}520\,17 \end{pmatrix} -$$

$$- 13\,912{,}2025 = 16\,795{,}0536.$$

Заполним клетки таблицы дисперсионного анализа 17.1.3 и вычислим критерий

$$F_{1,\,2} = \frac{16795{,}0536}{0{,}0870} \quad 1{,}93 \times 10^5,$$

который значим в высокой степени. Если модель

$$y_i = \beta_0 + \beta_1 (x_i \exp z_i)^{\frac{1}{2}} + e_i \qquad (17.1.4)$$

выбрана правильно, то имеются достаточные основания считать β_1 не равным нулю.

Для вычисления точечной оценки ожидаемого значения y в точке $x = 10$, $z = 2$ образуем вектор-строку

$$x_0 = (1;\ \sqrt{10e^2}) = (1;\ 8{,}5959619).$$

Заметим, что этот вектор имеет ту же размерность, что и строки матрицы X. Точечная оценка ожидаемого значения y при $x = 10$ и $z = 2$ дается выражением (15.10.2):

$$y_0 = \mathbf{x}_0\mathbf{b} = (1 \;\; 8{,}595\,961\,9) \begin{pmatrix} -0{,}058\,288\,883 \\ 3{,}000\,531\,915 \end{pmatrix} = 25{,}7342.$$

Для получения 95 %-ных доверительных границ для y при $x = 10$ и $2 = 2$ найдем

$$\mathbf{S}^{-1} = \begin{pmatrix} 0{,}457497354 & -0{,}010\,546\,633 \\ -0{,}010\,546\,633 & 0{,}000\,536\,062 \end{pmatrix}$$

[3] См. сноску 2.

к $\mathbf{x}_0\mathbf{S}^{-1}\mathbf{x}'_0 = 0{,}3158$. Средний квадрат остатков в таблице дисперсионного анализа 17.1.3 равен 0,0870. Регрессионное уравнение содержит два неизвестных параметра b, поэтому применим формулу (16.8.2) при $k+1 = 2$ и решим уравнения

$$(y - 25{,}7342)/\sqrt{0{,}0870} \times 0{,}3158 = \pm 4{,}303.$$

Доверительный интервал будет следующим: $25{,}73 \pm 0{,}71$.

Таблица 17.1.3. **Дисперсионный анализ. Пример множественной регрессии 17.1.2**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	16 795,0536	1	16 795,0536
Остаток	0,1739	2	0,0870
Общая вариация	16 795,2275	3	—

Таблица **17.1.4. Дисперсионный анализ. Пример множественной регрессии 17.1.3**

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	30 707,2488	1	30 707,2488
Остаток	0,1812	3	0,0604
Общая вариация	30 707,4300	4	—

Пример 17.1.3. С помощью метода наименьших квадратов подберем кривую вида

$$Y = b_1 (x \exp z)^{\frac{1}{2}} \qquad (17.1.5)$$

к четырем точкам из примера 17.1.1. Проверим нулевую гипотезу о том, что неизвестный коэффициент регрессии β_1 равен 2,7183.

В этом примере матрица X состоит из одного столбца, составленного из значений $\{\sqrt{x_i \exp z_i}\}$. Матрица S имеет единственный элемент $S = \sum x_i \exp z, = 3413{,}763\,833$.

Единственный элемент $\mathbf{X}'\mathbf{y}$ равен $\sum y_i (x_i \exp z_i)^{\frac{1}{2}} = 10\,238{,}520\,17$,

а нормальное уравнение (15.3.3) в данном случае имеет вид

$$3413{,}763\,833 b_1 = 10\,238{,}520\,17$$

с решением $b\backslash = 2{,}999\ 188\ 189$. Выравненными значениями y будут соответственно 8,565; 60,112; 2,999; 164,352. Уравнение (17.1.5) не содержит члена (общего среднего) b_0, поэтому дисперсионный анализ, представленный в табл. 17.1.4, является нескорректированным (см. табл. 16.4.2).

Единственный элемент обратной матрицы S^{-1} равен 0,000 292 93. Для проверки нулевой гипотезы $\beta_1 = 2{,}7183$ воспользуемся статистикой (16.6.1) и вычислим

$$t_3 = (2{,}999\ 188\ 189 - 2{,}7183)/\sqrt{0{,}0604 \times 0{,}000\ 292\ 93} = 66{,}8.$$

Эта величина значима в высокой степени, поэтому мы отклоняем нулевую гипотезу и делаем вывод, что $\beta_1 \neq 2{,}7183$.

Таблица 17.1.5. Значения зависимой переменной y, соответствующие различным значениям независимых переменных x и z (пример 17.1.4)

x	z = 1	z = 8	z = 32
1	23	52	112
2	31	63	105
5	28	73	125
10	50	67	144
50	112	167	197
100	213	261	308

Пример 17.1.4. С помощью метода наименьших квадратов подберем кривую вида

$$Y = b_1 x + b_2 \sqrt{z} \qquad (17.1.6)$$

к 18 точкам из табл. 17.1.5. Известно, что дисперсия зависимой переменной y в данной точке приблизительно пропорциональна математическому ожиданию y. Проверим нулевую гипотезу $\beta_1 = 2$.

Дисперсия в точке (x_i, z_i, y_i) приблизительно пропорциональна ожидаемому значению y_i, а в качестве аппроксимации этого ожидаемого значения может быть выбрано само значение уг. В качестве весов w_i в точке (x_i, z_i, y_i) возьмем $1/y_i$. Векторы и матрицы регрессии тогда будут иметь вид (см. параграф 16.10):

$$\mathbf{X}_w = \begin{pmatrix} x_1/\sqrt{y_1} & \sqrt{z_1}/\sqrt{y_1} \\ x_2/\sqrt{y_2} & \sqrt{z_2}/\sqrt{y_2} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{18}/\sqrt{y_{18}} & \sqrt{z_{18}}/\sqrt{y_{18}} \end{pmatrix}, \quad \mathbf{y}_w = \begin{pmatrix} y_1/\sqrt{y_1} \\ y_2/\sqrt{y_2} \\ \cdot \\ \cdot \\ \cdot \\ y_{18}/\sqrt{y_{18}} \end{pmatrix},$$

$$\mathbf{S}_w = \mathbf{X}_w'\mathbf{X}_w = \begin{pmatrix} \Sigma\, x_i^2/y_i & \Sigma\, (x_i/y)\sqrt{z_i} \\ \Sigma\, (x_i/y_i)/\sqrt{z_i} & \Sigma\, z_i/y_i \end{pmatrix} =$$

$$= \begin{pmatrix} 173{,}63759 & 8{,}1\,459\,947 \\ 8{,}145\,994\,7 & 2{,}0\,77\,4465 \end{pmatrix},$$

$$\mathbf{X}_w'\mathbf{y}_w = \begin{pmatrix} \sum x_i \\ \Sigma\, \sqrt{z_i} \end{pmatrix} \begin{pmatrix} 504 \\ 56{,}911\,688\,25 \end{pmatrix}$$

$$\mathbf{S}_w^{-1} = \begin{pmatrix} 0{,}00705737 & -0{,}02\,767\,305 \\ -0{,}02767305 & 0{,}58\,987\,057 \end{pmatrix}.$$

Нормальные уравнения (15.3.3) в данном примере имеют решения: $b_1 = 1{,}981\ 993\ 7$ и $b_2 = 19{,}623\ 311$.

Уравнение (17.1.6) не содержит члена (общего среднего) bo. При проведении регрессионного дисперсионного анализа в данном случае воспользуемся формулами нескорректированного дисперсионного анализа (см. табл. 16.4.2):

$$\text{общая с. к.} = \mathbf{y}_w'\mathbf{y}_w = y_1 + y_2 + \ldots + y_{18} = 2128,$$

$$\text{с. к. регрессии b'} (\mathbf{X}_w'\mathbf{y}_w) = 1{,}981\ 993\ 7 \times 504 + 19{,}623311 \times 59{,}911\,688\,25 = 2115{,}72.$$

Дисперсионный анализ представлен в табл. 17.1.6. Для проверки нулевой гипотезы $\beta_1 = 2$ воспользуемся статистикой (16.6.1) при $n = 18$, $\nu = 2$. Найдем

$$t_{16} = (1{,}981\ 993\ 7 - 2)/\sqrt{0{,}007\,057\,37 \times 0{,}768} = -0{,}245.$$

Эта величина незначима при 5 %-ном уровне, поэтому мы имеем право принять нулевую гипотезу.

Пример 17.1.5. Подберем кривую вида

$$Y = b_1 x + b_2 z \qquad (17.1.7)$$

к данным табл. 17.1.7. Проверим неадекватность подгонки.

Таблица 17.1.6. Дисперсионный анализ. Пример множественной регрессии 17.1.4

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	2115,72	2	1057,860
Остаток	12,28	16	0,768
Общая вариация	2128,00	18	—

Таблица 17.1.7. Данные для проверки неадекватности подгонки. Пример 17.1.5

Независимые переменные		Наблюдения зависимой переменной				Число наблюдений в точке (x_i, z_i)	Общее число наблюдений в точке (x_i, z_i)
x_i	z_i	$y_{i,j}$				n_i	T_i
1	2	2	5	3	2	4	12
3	4	15	14	12	15	4	56
5	4	34	30	36		3	100
9	6	70	66	64	67	4	267
Всего						15	435

Таблица 17.1.8. Дисперсионный анализ. Пример 17.1.5

Источник вариации		с. к.	с. с.	ср. к.
Регрессия		21 969,219	2	10 985
Остаток	Неадекватность подгонки	6,354	2	3,182
	Чистая ошибка	49,417	11	4,492
		55,781	13	4,291
Общая вариация		22 025,000	15	—

Нормальными уравнениями будут:

$$\begin{pmatrix} 439 & 332 \\ 332 & 272 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 3083 \\ 2250 \end{pmatrix},$$

откуда $b_1 = 9{,}971\,2544$ и $b_2 = -3{,}898\,736\,9$. Уравнение (17.1.7) не имеет коэффициента общего среднего b_0, поэтому мы применяем нескорректированный дисперсионный анализ неадекватности подгонки (см. табл. 16.4.2 и сноску 7 в параграфе 16.9). Вычисления приводят нас к следующим значениям:

$$\text{общая с. к.} = 2^2 + 5^2 + 3^2 + \ldots + (67)^2 = 22\,025,$$

$$\text{с. к. регрессии} = 9{,}971\,254\,4 \times 3\,083 - 3{,}898\,736\,9 \times 2\,250 = 21\,969{,}219,$$

$$\text{остаточная с. к.} = 22\,025 - 21\,969{,}219 = 55{,}781,$$

$$\text{с. к. регрессии} + \text{с. к. неадекватности} = (12)^2/4 + (56)^2/4 + (100)^2/3 + (267)^2/4 = 21\,975{,}583,$$

$$\text{с. к. неадекватности} = 21\,975{,}583 - 21\,969{,}219 = 6{,}364,$$

$$\text{с. к. чистой ошибки} = 55{,}781 - 6{,}364 = 49{,}417.$$

Дисперсионный анализ приведен в табл. 17.1.8. Для проверки неадекватности подгонки найдем

$$F_{2,11} = 3{,}182/4{,}492 = 0{,}71.$$

Это значение незначимо, поэтому у нас нет свидетельств неадекватности подгонки.

Литература: [7, с. 419—465], [19, с. 104—127; русский перевод с 117—133], [45, с. 207—208], [46, с. 174—176], [50, с. 413—422; русский перевод с. 487—498], [66, с. 135], [105, с. 252—280].

17.2. КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Большинство программ регрессионного анализа на ЭВМ предусматривает выдачу на печать значения *коэффициента детерминации*. Этот коэффициент представляет собой объясненную регрессией долю общей суммы квадратов, т. е. является отношением суммы квадратов регрессии к общей сумме квадратов в дисперсионном анализе. Этот показатель лежит в пределах от нуля до единицы. При скорректированном дисперсионном анализе (см. табл. 16.4.1) коэффициент детерминации равен квадрату коэффициента корреляции между наблюдаемыми расчетными значениями y. Последний называется *коэффициентом множественной корреляции*. Коэффициент детерминации обозначается как R^2.

Коэффициент детерминации иногда используется в качестве критерия при выборе «наилучшего» уравнения регрессии (см. параграф 17.3).

Пример 17.2.1. Коэффициент детерминации в примере 17.1.4 равен $2115{,}72/2128{,}00 = 0{,}9942$.

Литература: [19, с. 26; русский перевод с. 34], [105, с. 260].

17.3. КАКУЮ МОДЕЛЬ ИСПОЛЬЗОВАТЬ?

К данным из примера 17.1.1 было подобрано три различных регрессионных уравнения (17.1.1), (17.1.3) и (17.1.5), причем в каждом случае регрессия оказывалась высокозначимой. Какую же регрессионную модель выбрать?

При ответе на этот вопрос необходимо принять во внимание следующие моменты:

1) могут быть основательные теоретические причины, которые заставляют нас поверить в то, что регрессионное уравнение имеет определенную математическую форму;

2) регрессионное уравнение должно обеспечить наибольшее приближение к наблюдаемым значениям (тогда уравнение может быть с достаточной уверенностью использовано в прогностических целях);

3) регрессионное уравнение должно быть как можно более простым.

Требования 2 и 3 часто противоречат друг другу, поэтому необходим компромисс. Окончательный выбор *«наилучшего» регрессионного уравнения,* таким образом, становится субъективным.

Уравнение (17.1.5) представляет собой частный случай (17.1.3), поэтому в действительности нам необходимо сравнить только два основных уравнения (17.1.1) и (17.1.3). У нас нет информации о теоретическом обосновании этих уравнений, которая указывала бы на предпочтительную математическую форму. Аналогично трудно выбрать уравнение, которое было бы более простым по сравнению с другим. Единственное, что можно утверждать, это то, что уравнение (17.1.1) обеспечивает лучшую подгонку (меньшую остаточную сумму квадратов). Поэтому выберем уравнение (17.1.1).

Уравнение (17.1.1) содержит три параметра b_0, b_1 *и* b_2. Возможно, один (или более) параметр из β_0, β_1 и β_2 весьма мал или равен нулю. Если такой член опустить из уравнения, остаточная сумма квадратов ненамного увеличится, а модель станет проще. Коэффициенты регрессии и остаточные суммы квадратов всех редуцированных уравнений (17.1.1) показаны в табл. 17.3.1. При этом была использована нескорректированная[4] сумма квадратов, поскольку мы хотим изучить также и возможность исключения *bo*. Как видим, уравнение (4) с двумя параметрами обеспечивает практически то же качество подгонки, что и трехпараметрическое уравнение (1), однопараметрическое уравнение (7) также достаточно удовлетворительно.

Был предложен ряд объективных методов выявления «наилучших» уравнений регрессии. Они не обязательно приводят к одинаковым результатам. Остаточный средний квадрат оценивает дисперсию σ^2, и один из объективных методов заключается в выборе уравнения с наименьшим средним квадратом остатка (уравнение (4) в табл. 17.3.1). Такой подход требует вычисления остаточного среднего квадрата для всех возможных редуцированных регрессий (если начальная модель содержит v параметров, то общее число регрессий равно 2^{v-1}). Общее количество вычислений велико даже для сравнительно небольших v (при $v = 10$ общее число регрессий равно 1023).

Таблица 17.3.1. **Альтернативные уравнения регрессии для данных из примера 17.1.1**

Регрессионное уравнение	Нескорректированная сумма квадратов регрессии	Остаточная сумма квадратов	Остаточный средний квадрат
(1) $b_0 + b_1x + b_2e^z$	30707,4134	0,0166	0,0166
(2) $b_0 + b_1x$	30706,5936	0,8364	0,4182
(3) $b_0 + b_2e^z$	30707,4037	0,0263	0,0132
(4) $b_1x + b_2e^z$	30707,4133	0,0167	0,0084
(5) b_0	13912,2025	16795,2275	5598,4091
(6) b_1x	30706,5526	0,8774	0,2925
(7) b_2e^z	30707,4032	0,0268	0,0089

[4] См. табл. 16.4.2.

Было предложено большое число последовательных процедур по выбору «наилучшего» регрессионного уравнения без расчета всех возможных вариантов. Рассмотрим, например, известную *процедуру исключения*.

Допустим, к данным была подобрана v-параметрическая модель. Для проверки нулевой гипотезы $\beta_i = 0$ для каждого i можно найти значение F-статистики (16.7.1) или, эквивалентно, t-статистики (16.6.1). Более простая модель из $v - 1$ параметров может быть получена из предыдущей в результате исключения члена с наименьшим значением F при условии, что значение F незначимо. Процесс повторяется для $v - 1$-параметрической модели, на основе которой получают модель с $v - 2$ параметрами, и т. д. Процедура заканчивается, когда все параметры будут значимыми. Если процесс начать с модели, содержащей v параметров, а конечная модель содержит p параметров, то число необходимых регрессий равно $(v(v+1) - p(p-1))/2$ (если $v = 10$ и $p = 3$, то необходимо рассчитать 52 регрессии). Иногда применяют и другие методы усечения регрессий.

Если применить процедуру исключения к трехпараметрической модели из табл. 17.3.1, то мы придем к уравнению (7) как к «наилучшему». Вычисления сведены в табл. 17.3.2.

Таблица 17.3.2. **Процедура исключения**

Регрессионное уравнение	Остаточный средний квадрат	Исключаемый член	Увеличение в остаточной сумме квадратов	F
$b_0 + b_1x + b_2e^z$	0,0166	b_2e^z	0,8198	$F_{1,1} = 49,4$
		b_1x	0,0097	$F_{1,1} = 0,6$
		b_0	0,0001	$F_{1,1} = 0,006$
$b_1x + b_2e^z$	0,0084	b_2e^z	0,8607	$F_{1,2} = 102,5$
		b_1x	0,0101	$F_{1,2} = 1,2$

Процедура исключения часто приводит к удовлетворительным результатам, однако и она не идеальна. Дрейпер и Смит [19, с. 163—177; русский перевод с. 177—178] описывают некоторые альтернативные процедуры, которые затем сравниваются (см. также [44]). Среди простейших процедур выбора «наилучшей» регрессии метод исключения, вероятно, наиболее удовлетворительный. Читателю необходимо помнить о предупреждении, сделанном в параграфе 12.2, относительно применения нескольких критериев к одному и тому же множеству данных.

В приведенном примере взята нескорректированная сумма квадратов (см. табл. 16.4.2), поскольку мы хотели исследовать ситуацию, когда член в b_0 также исключается из регрессии. Если бы с самого начала предполагалось, что член общего среднего автома-

тически должен присутствовать в уравнении, необходимо было бы взять скорректированную сумму квадратов (см. табл. 16.4.1).

В нашем примере e^z и x сильно коррелируют. Неудивительно поэтому, что удовлетворительной подгонки можно добиться включением в регрессионное уравнение лишь одной из этих переменных. Высокая корреляция порождает определенные проблемы при решении нормальных уравнений регрессий (1) и (4) из табл. 17.3.1, поскольку матрицы S в таком случае становятся почти вырожденными (см. параграф 1.8).

Литература: [19, с. 163—195; русский перевод с. 115—133], [44], [105, с. 263—266].

17.4. УПРАЖНЕНИЯ

1. Найдите сумму квадратов разностей между наблюдаемыми и расчетными значениями y в примерах 17.1.1 и 17.1.2. Проверьте, что эти суммы равны остаточным суммам квадратов.

2. Найдите матрицы X, S и вектор X'y при подгонке методом наименьших квадратов следующих функций:

а) $Y = b_0 + b_1x + b_2x^2 + b_3z$;

б) $Y = b_0 + b_1x + b_2z + b_3w$;

в) $Y = b_0 + b_1 \sin \pi x + b_2 \cos \pi z$.

3. Подберите функцию вида $Y=b_0+b_1x+b_2z$ к данным табл. 17.4.1 и постройте таблицу дисперсионного анализа.

Таблица 17.4.1. Данные для упражнения 3 из параграфа 17.4

x_i	z_i	y_i	x_i	z_i	y/
1	1	2,1	6	6	11,9
2	8	9,9	7	7	8,9
3	4	6,9	8	8	15,1
4	9	13,1	9	9	11,9
5	5	10,1	10	10	20,1

Таблица 17.4.2. Данные для упражнения 4

Объем легких, л x_i	Число пачек сигарет, выкуренных за неделю z_i	Оценка y_i	Объем легких, л $*i$	Число пачек сигарет, выкуренных за неделю z_i	Оценка y_i
5,0	4	483	5,5	0	556
4,5	10	351	5,4	4	510
4,6	6	390	5,2	2	509
4,8	4	451	4,6	3	440
4,7	2	439	5,0	3	480

4. В табл. 17.4.2 приводятся оценки $\{y_i\}$, полученные десятью курсантами после окончания курсов по прыжкам в воду. В этой же таблице показаны объем их легких $\{x_i\}$ и количество пачек сигарет, выкуренных ими за неделю. Подберите к этим данным регрессию вида $Y=b_0+b_1x+b_2z$ и постройте таблицу дисперсионного анализа. Необходимо ли в данном случае наличие двух независимых переменных x и z?

18. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Матричные методы, рассмотренные в гл. 15—17, при переходе к нелинейной регрессии перестают играть свою роль. Однако, как будет показано в этой главе, метод наименьших квадратов применим и в этом случае. Статистический анализ теперь будет приближенным.

18.1. ВВЕДЕНИЕ. НЕКОТОРЫЕ ПРИМЕРЫ НЕЛИНЕЙНЫХ РЕГРЕССИЙ

Матричные методы, описанные в гл. 15—17, применимы тогда, когда регрессионное уравнение линейно по коэффициентам $\{b_i\}$. Однако они теряют силу, как только модель перестает быть «линейной». Так, эти методы пригодны при подгонке функции вида

$$Y = b_0 + b_1 \sin(\pi e^x) + b_2 \ln z,$$

однако их нельзя применять при выравнивании по уже достаточно простой нелинейной функции [1] вида

$$y = b_0 + b_1 \exp(b_2 x). \qquad (18.1.1)$$

Хотя обычные матричные методы перестают быть полезными, принцип метода наименьших квадратов остается по-прежнему приемлемым. Нормальные уравнения становятся нелинейными [2], и для их решения необходимо применение итеративных методов (например, многомерный метод Ньютона—Рафсона из параграфа 3.6). Иногда решение, основанное на методе наименьших квадратов, может быть найдено методом наискорейшего спуска (см. параграф 7.5). Очевидно, что для получения нелинейных регрессий необходима помощь ЭВМ.

Статистические методы линейной регрессии могут быть адаптированы для нелинейной регрессии, однако статистический анализ при этом будет приближенным. Общая сумма квадратов остатков в дисперсионном анализе вычисляется обычным способом, на ЭВМ также нетрудно подсчитать остаточную сумму квадратов как сумму квадратов разностей между наблюдаемыми и расчетными значениями y. Сумма квадратов регрессии тогда может быть

[1] Исследователи часто линеаризуют (18.1.1), прежде чем приступить к регрессионному анализу. Уравнение может быть переписано в следующем виде: $\ln(Y-b_0)=\ln\&1+b_2x$, оно приводится к линейной регрессии $\ln(Y-b_0)$ на x. Для этого необходимо выбрать подходящее значение b_0 (часто можно положить $b_0=0$); $\ln b_1$ и b_2 тогда оцениваются с помощью парной линейной регрессии (более подробно об этом см. в примере 18.1.2).

[2] В обычном математическом смысле.

найдена как разность. По аналогии с линейной регрессией каждому расчетному значению переменной соответствует одна степень свободы.

Пример 18.1.1. С помощью метода наименьших квадратов подберем кривую вида (18.1.1) к четырем точкам:

$$(x_1,\ y_1) = (1;\ 5{,}4), \qquad (x_3,\ y_3) = (4;\ 26{,}2),$$
$$(x_2,\ y_2) = (3;\ 15{,}3), \qquad (x_4,\ y_4) = (5;\ 47{,}4).$$

Нам необходимо минимизировать

$$\sum_{i=1}^{4} \{y_i - b_0 - b_1 \exp(b_2 x_i)\}^2.$$

В точке минимума частные производные по b_0, b_1 и b_2 равны нулю (см. параграф 1.6). Таким образом, нелинейными нормальными уравнениями являются:

$$4b_0 + b_1 \sum \exp(b_2 x_i) = \sum y_i,$$
$$b_0 \sum \exp(b_2 x_i) + b_1 \sum \exp(2b_2 x_i) = \sum y_i \exp(b_2 x_i),$$
$$b_0 \sum x_i \exp(b_2 x_i) + b_1 \sum x_i \exp(2b_2 x_i) = \sum y_i x_i \exp(b_2 x_i),$$

решить которые вручную довольно трудно, однако на ЭВМ это можно сделать довольно быстро.

В данном случае возможен и другой подход. Допустим, что параметр b_2 известен и необходимо найти только b_0 и b_1. Зависимость (18.1.1) является линейной по неизвестным параметрам b_0, b_1 и для их определения может быть применен обычный матричный подход. В практике, разумеется, b_2 не является известным и можно проделать несколько подгонок, каждый раз применяя линейный метод наименьших квадратов. Значение b_2 и соответствующие значения b_0 и b_1, отвечающие минимальной сумме квадратов остатков, приведут нас к кривой с наименьшими квадратами, наилучшим образом подогнанной к точкам. Проиллюстрируем этот метод.

Прежде всего необходимо выбрать начальное значение b_2. Если через $Y(x)$ обозначить расчетное значение y в точке x, то, как легко видеть из (18.1.1),

$$\ln\left(\frac{Y(5) - Y(4)}{Y(4) - Y(3)}\right) = b_2 = \frac{1}{2} \ln\left(\frac{Y(5) - Y(3)}{Y(3) - Y(1)}\right).$$

Таким образом, подставляя в эту формулу вместо расчетных значений наблюдаемые значения y, получим два различных приближения для b_2: 0,588 16 и 0,66524. В качестве начального значения b_2 выберем 0,63.

Для вычисления значений b_0 и b_1, соответствующих данному значению $b_2 = 0{,}63$, воспользуемся матричными методами линейной регрессии. Первый столбец регрессионной матрицы X будет тогда состоять из единиц, второй — из значений $\{\exp(0{,}63x_i)\}$.

Значения $\{y_i\}$, как и ранее, образуют вектор-столбец y. Нормальные уравнения для b_0 и b_1 будут иметь следующий вид:

$$4b_0 + 44{,}261\ 640\ 46 b_1 = 93{,}4,$$
$$44{,}261\ 640\ 46 b_0 + 746{,}383\ 387\ 1 b_1 = 1543{,}174\ 130.$$

Система уравнений имеет решение: $b_0 = 2{,}026\ 941\ 082$ и $b_1 = 1{,}947\ 334\ 866$.

Поскольку в зависимости (18.1.1) значение b_2 было задано, сумма квадратов регрессии должна быть нескорректированной[3]; она находится стандартным способом по формуле

$$\mathbf{b'X'y} = (b_0\ \ b_1)\begin{pmatrix} 94{,}3 \\ 1543{,}174\ 130 \end{pmatrix} = 3196{,}217\ 331.$$

Необходимо отыскать такое значение b_2, которое минимизирует сумму квадратов остатков (или максимизирует сумму квадратов регрессии, что эквивалентно). Применим метод проб и ошибок (варьируя значения параметра b_2).

Последовательные приближения по этому методу показаны в табл. 18.1.1. Искомым решением будет:

$$b_0 = 1{,}471\ 055, \quad b_1 = 2{,}183\ 988, \quad b_2 = 0{,}608\ 957.$$

(Для получения более точных приближений параметра b_2 в табл. 18.1.1 была использована формула квадратичного минимума (7.5.2).)

Таблица 18.1.1. **Последовательные приближения к кривой с наименьшими квадратами остатков**

b_2	b_0	b_1	с. к. регрессии
0,63	2,026941082	1,947334866	3196,217331
0,62	2,056273539	1,767252438	3196,294192
0,60	1,223215224	2,293638769	3196,304055
0,58	0,643464106	2,559808774	3196,114497
0,61	1,499478421	2,171574427	3196,323481
0,608987683	1,471904985	2,183616581	3196,323750
0,607975366	1,444244145	2,195728454	3196,323518
0,609025063	1,472924743	2,183170672	3196,323746
0,608950303	1,470885177	2,184062571	3196,323751
0,608956533	1,471055142	2,183988233	3196,323751

Хотя регрессионная модель и не является линейной, для приближенного статистического анализа может быть построена таб-

[3] Т. е. без поправки $n\bar{y}^2$ (см. параграф 16.4).

лица дисперсионного анализа. Статистическая модель в данном примере имеет вид:

$$y_i = \beta_0 + \beta_1 \exp(\beta_2 x_i) + e_i. \qquad (18.1.2)$$

Значения $\{x_i\}$ не содержат ошибки, $\{e_i\}$ — независимые случайные переменные, распределенные по нормальному закону с нулевым математическим ожиданием и одинаковой дисперсией σ^2. В зависимости от целей исследования дисперсионный анализ может быть скорректированным (см. табл. 16.4.1) или нескорректированным (см. табл. 16.4.2) с $n = 4$ и $v = 3$. Остаточная сумма квадратов, разумеется, должна быть одинаковой в обоих случаях.

В целях проверки значимости члена $b_1 \exp(b_2 z)$ дисперсионный анализ должен быть скорректированным (см. табл. 18.1.2). Общая сумма квадратов была подсчитана по формуле из табл. 16.4.1, а сумма квадратов регрессии получена вычитанием $n\bar{y}^2$ из нескорректированной суммы квадратов регрессии (последняя строка табл. 18.1.1). Отношение средних квадратов равно:

$$486{,}6006/0{,}1262 = 3856,$$

оно намного превышает 5 %-ную точку $F_{2,1}$-распределения. Весьма вероятно, что экспоненциальный член объясняет значимую пропорцию вариации переменной y,

Таблица 18.1.2. Дисперсионный анализ. Нелинейный метод наименьших квадратов

Источник вариации	с. к.	с. с.	ср. к.
Регрессия	973,201251	2	486,6006
Остаток	0,126249	1	0,1262
Общая вариация	973,327500	3	—

Пример 18.1.2. Метод линеаризации, часто применяемый исследователями при выравнивании данных по кривой вида (18.1.1), был описан в сноске 1 на с. 319. Применим его для данных примера 18.1.1.

Усилия, направленные на определение b_0, зависят от желательной точности и целей применения уравнения регрессии. Параметры b_1 и b_2 могут компенсировать неточно найденное значение b_0, поэтому часто бывает достаточно грубой оценки этого параметра. Так, например, из табл. 18.1.1 видно, как незначительно меняется сумма квадратов регрессии при изменении значений параметра b_2 и насколько заметно при этом изменение соответствующих значений параметров b_0 и b_1. Если требуется высокая точность, то значения $\ln(y_i - b_0)$ вместе с x_i могут быть нанесены на график для различных значений b_0, после чего необходимо выбрать то значение b_0, диаграмма рассеяния которого будет наиболее близка к линейной.

В нашем примере выбор $bo = 0$ не кажется неудовлетворительным. Если применить методы парной линейной регрессии (см. гл. 15) к регрессии $\ln y$ на x, то мы получим уравнение

$$\ln Y = 1{,}129\,455 + 0{,}540\,062\,6x$$

или

$$Y = 3{,}093\,97e^{0{,}540\,062\,6x}.$$

Расчетные значения равны соответственно 5,3; 15,6; 26,8 и 46,1. Необходимо отметить, что расчетные значения в точках $x = 1$ и $x = 5$ лежат ниже наблюдаемых значений y, а для $x = 2, 3, 4$ расчетные значения превышают наблюдаемые. Лучшей подгонки можно добиться, выбирая bo несколько больше нуля; кривая регрессии будет тогда более крутая (параметр b_2 увеличится).

Статистическая модель, лежащая в основе метода линеаризации, имеет вид:

$$\ln(y_i - \text{Po}) = \ln \beta_1 + faXi + e_i.$$

Значения $\{x_i\}$ свободны от ошибок, $\{e_i\}$ — независимые случайные переменные, распределенные по нормальному закону с нулевым математическим ожиданием и одинаковой дисперсией σ^2. Эта модель не совпадает с моделью (18.1.2). Разные модели приводят, в частности, к разным коэффициентам b, поскольку они дают разные веса различным наблюдениям.

Пример 18.1.3. В табл. 18.1.3 представлены данные о влиянии температуры на относительную подвижность жирных кислот в хлоропластных оболочках. В соответствии с теоретическими соображениями $\log_{10}\tau$ (логарифм относительной подвижности т) и величина, обратная к абсолютной температуре, линейно зависимы. Из теории известно также, что жирные кислоты при некоторой температуре начинают плавиться, поэтому в точке плавления наклон прямой меняется.

Для обнаружения точки плавления жировой кислоты методом наименьших квадратов оценим пару прямых линий, проходящих близко к данным. При этом может быть предложен следующий подход.

С помощью метода из параграфа 15.2 подберем прямую к первым n точкам, вычислим для этих точек остаточную сумму квадратов по формуле из табл. 15.5.1. Построим другую прямую по $23 - n$ точкам и также вычислим остаточную сумму квадратов для пары прямых сложением соответствующих остаточных сумм квадратов. Выполним эти вычисления для каждого га от 2 до 21 включительно. В качестве окончательного результата выберем ту пару прямых, которая дает наименьшую остаточную сумму квадратов.

Численные значения остаточных сумм квадратов для различных разбиений 23 точек приведены в табл. 18.1.4. Как видим, пара прямых, соответствующих методу наименьших квадратов, получается при выравнивании первых 12 наблюдений по одной прямой,

Таблица 18.1.3. **Влияние температуры на сравнительную подвижность жирных кислот в хлоропластных оболочках**

Температура, °С	10^4 К$^{-1}$	Сравнительная подвижность т	Температура, °С	10^4 К$^{-1}$	Сравнительная подвижность т
0,1	3 661	63,2	14,3	3 480	24,3
1,6	3 641	54,1	15,9	3 451	23,6
3,1	3 621	49,7	17,4	3 443	21,6
4,7	3601	42,5	19,0	3 424	19,6
6,1	3 582	40,0	20,7	3 404	19,1
7,6	3 563	37,5	22,3	3 386	17,5
9,0	3 546	33,1	23,6	3 371	16,7
10,0	3533	31,0	25,1	3 354	16,6
11,1	3519	30,3	26,8	3 335	15,8
11,9	3510	28,0	28,3	3318	14,5
12,7	3 500	26,2	29,8	3 302	13,7
13,5	3 490	25,3			

Источник. Сообщено в частной беседе Дж. К. Рейзеном, Научно-промышленный исследовательский центр, отдел по изучению питания. Норс Рид, Австралия.

Таблица 18.1.4. Нахождение пары прямых методом наименьших квадратов. Остаточная сумма квадратов для различных разбиений 23 точек; n обозначает число точек, использованных для получения первой прямой

n	(остаточная с к)× X 10^6	n	(остаточная с. к.) X X 10^6
3	6711	11	2 250
4	6 197	12	2 243 *
5	5 300	13	2 285
6	3 879	14	2915
7	3 426	15	3410
8	3 166	16	3 659
9	2 507	17	5 220
10	2 245	18	6061

* Соответствует искомой паре прямых по методу наименьших квадратов.

а остальных 11 наблюдений — по другой. Обозначим через Y расчетное значение $\log_{10}\tau$, а через x — число 10^4, умноженное на величину, обратную к абсолютной температуре. Тогда уравнения прямых можно записать так:

$$Y = 0{,}002\,242x - 6{,}424\,576 \text{ и } Y = 0{,}001\,380x - 3{,}417718.$$

Эти прямые изображены на рис. 18.1.1. Разница в наклоне прямых в точке плавления не является существенной. Возможно, одной прямой можно добиться также достаточно адекватной подгонки. В параграфе 16.6 был описан F-метод проверки того, влечет ли введение нового члена в уравнение регрессии значимое уменьшение суммы квадратов. Нескорректированный дисперсионный анализ показан в табл. 16.7.1. Проведем скорректированный дисперсионный анализ для нашей кусочно-линейной модели.

Сумма квадратов регрессии одного уравнения, как легко показать, равна 0,776602, а общая сумма квадратов равна 0,792 229

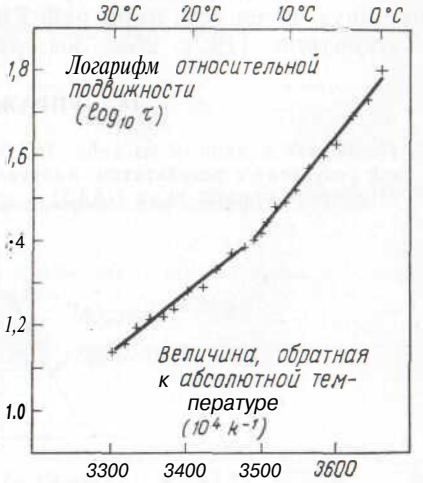


Рис. 18.1.1. Влияние температуры на относительную подвижность жирных кислот в хлоропластных оболочках. Пара прямых

(формулы для вычислений приведены в табл. 15.5.1). Сумма квадратов остатков при условии, что подгоняется пара прямых, равна 0,002 243 (см. табл. 18.1.4). Уменьшение суммы квадратов, связанное с выделением точки плавления и переходом к паре прямых, находится как разность. Эти суммы квадратов занесены в табл. 18.1.5. Степени свободы получают исходя из тех соображений, что прямая линия содержит две неизвестные константы, а пара прямых вместе с неизвестной точкой плавления — приблизительно пять.

Таблица 18.1.5. **Дисперсионный анализ. Кусочно-линейная регрессия**

Источник вариации	с. к.	с. с.	ср. к.
Одна прямая линия	0,776602	1	—
Уменьшение, связанное с переходом к двум прямым	0,013384	3	0,0044613
Остаток	0,002243	18	0,0001246
Общая вариация	0,792229		—

Отношение средних квадратов велико (оно равно 35,8). Это означает, что уменьшение суммы квадратов, при учете точки плавления и переходе к паре прямых, значимо. Однако здесь мы не можем указать для F-критерия определенный уровень доверия.

Необходимо отметить, что приведенные вычисления были выполнены на программируемом калькуляторе, имеющем 51 ячейку памяти и допускающем 512 программных шагов. Рис. 18.1.1 был построен ЭВМ. Трудности, связанные с ограниченным размером памяти калькулятора, были преодолены при запоминании в одной ячейке двух чисел (см. параграф 7.6).

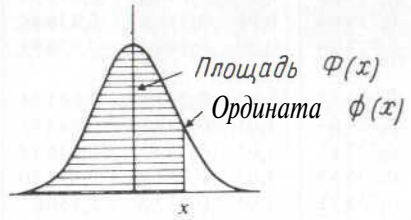
Литература: [19, с. 263—301; русский перевод с. 269—306].

18.2. УПРАЖНЕНИЯ

1. Подберите к данным из табл. 18.1.3 полином невысокого порядка и сравните свой результат с результатом, полученным в примере 18.1.3.
2. Подберите кривую вида (18.1.1) к данным из табл. 4.2.1.

ПРИЛОЖЕНИЕ

Таблица АЛ. Ординаты и площадь под нормальной кривой



Замечание: $\phi(-x)=\phi(x)$, $\Phi(-x)=1-\Phi(x)$

x	$\phi(x)$	$\Phi(x)$	x	$\phi(x)$	$\Phi(x)$	x	$\phi(x)$	$\Phi(x)$
0,00	0,398 94	0,500 00	0,15	0,39448	0,559 62	0,30	0,381 39	0,61791
0,01	0,39892	0,503 99	0,16	0,393 87	0,563 56	0,31	0,380 23	0,621 72
0,02	0,398 86	0,507 98	0,17	0,393 22	0,567 49	0,32	0,379 03	0,625 52
0,03	0,398 76	0,511 97	0,18	0,392 53	0,571 42	0,33	0,377 80	0,629 30
0,04	0,398 62	0,515 95	0,19	0,391 81	0,575 35	0,34	0,376 54	0,633 07
0,05	0,39844	0,51994	0,20	0,391 04	0,579 26	0,35	0,375 24	0,636 83
0,06	0,398 22	0,523 92	0,21	0,390 24	0,583 17	0,36	0,37391	0,64058
0,07	0,397 97	0,527 90	0,22	0,38940	0,587 06	0,37	0,37255	0,644 31
0,08	0,397 67	0,531 88	0,23	0,388 53	0,590 95	0,38	0,371 15	0,648 03
0,09	0,397 33	0,535 86	0,24	0,387 62	0,594 83	0,39	0,369 73	0,651 73
0,10	0,39695	0,53983	0,25	0,386 67	0,598 71	0,40	0,368 27	0,655 42
0,11	0,396 54	0,543 80	0,26	0,385 68	0,602 57	0,41	0,366 78	0,659 10
0,12	0,396 08	0,547 76	0,27	0,384 66	0,606 42	0,42	0,365 26	0,662 76
0,13	0,395 59	0,551 72	0,28	0,383 61	0,610 26	0,43	0,363 71	0,666 40
0,14	0,395 05	0,555 67	0,29	0,38251	0,614 09	0,44	0,362 13	0,67003

Источник. Kenney. Mathematics of Statistics. *Van* Nostrand. New York, Part one, p. 225—227. Воспроизведено с разрешения издателя.

Продолжение табл. АЛ

x	$\phi(x)$	$\Phi(x)$	x	$\phi(x)$	$\Phi(x)$	x	$\phi(x)$	$\Phi(x)$
0,45	0,36053	0,673 64	0,85	0,27798	0,80234	1,25	0,18265	0,89435
0,46	0,35889	0,67724	0,86	0,27562	0,805 11	1,26	0,180 37	0,89617
0,47	0,35723	0,68082	0,87	0,27324	0,807 85	1,27	0,17810	0,89796
0,48	0,35553	0,684 39	0,88	0,27086	0,81057	1,28	0,17585	0,89973
0,49	0,35381	0,68793	0,89.	0,26848	0,81327	1,29	0,17360	0,90147
0,50	0,35207	0,69146	0,90	0,26609	0,81594	1,30	0,17137	0,90320
0,51	0,35029	0,69497	0,91	0,26369	0,81859	1,31	0,169 15	0,90490
0,52	0,348 49	0,69847	0,92	0,26129	0,82121	1,32	0,16694	0,90658
0,53	0,34667	0,70194	0,93	0,25888	0,82381	1,33	0,164 74	0,90824
0,54	0,34482	0,705 40	0,94	0,25647	0,82639	1,34	0,16256	0,90988
0,55	0,342 94	0,708 84	0,95	0,25406	0,82894	1,35	0,16038	0,91149
0,56	0,341 05	0,71226	0,96	0,251 64	0,83147	1,36	0,15822	0,91309
0,57	0,339 12	0,71566	0,97	0,24923	.0,83398	1,37	0,15608	0,91466
0,58	0,337 18	0,71904	0,98	0,24681	0,83646	1,38	0,15395	0,91621
0,59	0,335 21	0,72240	0,99	0,24439	0,838 91	1,39	0,15183,	0,91774
0,60	0,333 22	0,725 75	1,00	0,241 97	0,841 34	1,40	0,14973	0,91924
0,61	0,331 21	0,72907	1,01	0,23955	0,84375	1,41	0,147 64	0,92073
0,62	0,329 18	0,73237	1,02	0,237 13	0,846 14	1,42	0,14556	0,92220
0,63	0,32713	0,73565	1,03	0,234 71	0,84850	1,43	0,14350	0,92364
0,64	0,32506	0,738 91	1,04	0,232 30	0,85083	1,44	0,14146	0,92507
0,65	0,32297	0,742 15	1,05	0,22988	0,853 14	1,45	0,139 43	0,92647
0,66	0,320 86	0,74537	1,06	0,22747	0,85543	1,46	0,13742	0,927 86
0,67	0,31874	0,74857	1,07	0,22506	0,85769	1,47	0,13542	0,92922
0,68	0,316 59	0,751 75	1,08	0,22265	0,859 93	1,48	0,13344	0,93056
0,69	0,31443	0,754 90	1,09	0,22025	0,862 14	1,49	0,13147	0,93189
0,70	0,312 25	0,75804	1,10	0,21785	0,86433	1,50	0,12952	0,93319
0,71	0,310 06	0,761 15	1,11	0,21546	0,86650	1,51	0,12758	0,93448
0,72	0,30785	0,764 24	1,12	0,213 07	0,86864	1,52	0,12566	0,93574
0,73	0,305 63	0,767 30	1,13	0,210 69	0,87076	1,53	0,12376	0,93699
0,74	0,303 39	0,77035	1,14	0,208 31	0,87286	1,54	0,12188	0,93822
0,75	0,301 14	0,77337	1,15	0,20594	0,87493	1,55	0,12001	0,93943
0,76	0,29887	0,77637	1,16	0,20357	0,876 98	1,56	0,11816	0,94062
0,77	0,29659	0,779 35	1,17	0,201 21	0,879 00	1,57	0,11632	0,941 79
0,78	0,29431	0,78230	1,18	0,198 86	0,88100	1,58	0,11450	0,94295
0,79	0,29200	0,78524	1,1.9	0,196 52	0,88298	1,59	0,112 70	0,94408
0,80	0,28969	0,788 14	1,20	0,19419	0,88493	1,60	0,11092	0,94520
0,81	0,28737	0,79103	1,21	0,19186	0,88686	1,61	0,10915	0,94630
0,82	0,28504	0,79389	1,22	0,18954	0,88877	1,62	0,10741	0,94738
0,83	0,28269	0,79673	1,23	0,18724	0,89065	1,63	0,10567	0,94845
0,84	0,280 34	0,79955	1,24	0,18494	0,89251	1,64	0,103 96	0,949 50

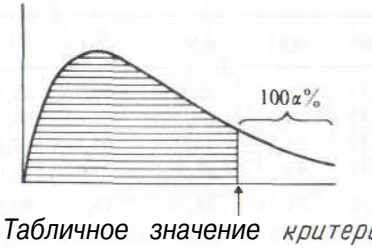
Продолжение табл. АЛ

x	$\phi(x)$	$\Phi(x)$	x	$\Phi(x)$	$\Phi(x)$	x	$\phi(x)$	$\Phi(x)$
1,65	0,10226	0,950 53	2,05	0,048 79	0,979 82	2,45	.0,01984	0,99286
1,66	0,10059	0,95154	2,06	0,047 80	0,98030	2,46	0,01936	0,99305
1,67	0,09893	0,952 54	2,07	0,046 82	0,98077	2,47	0,01889	0,993 24
1,68	0,09728	0,95352	2,08	0,04586	0,98124	2,48	0,01842	0,993 43
1,69	0,095 66	0,95449	2,09	0,04491	0,98169	2,49	0,01797	0,99361
1,70	0,09405	0,95543	2,10	0,043 98	0,98214	2,50	0,01753	0,99379
1,71	0,09246	0,95637	2,11	0,04307	0,98257	2,51	0,01709	0,99396
1,72	0,09089	0,95728	2,12	0,04217	0,98300	2,52	0,01667	0,994 13
1,73	0,08933	0,95818	2,13	0,041 28	0,98341	2,53	0,01625	0,99430
1,74	0,087 80	0,95907	2,14	0,04041	0,98382	2,54	0,015 85	0,99446
1,75	0,08628	0,95994	2,15	0,03955	0,98422	2,55	0,01545	0,99461
1,76	0,08478	0,96080	2,16	0,03871	0,98461	2,56	0,01506	0,99477
1,77	0,08329	0,96164	2,17	0,03788	0,98500	2,57	0,014 68	0,994 92
1,78	0,081 83	0,96246	2,18	0,03706	0,98537	2,58	.0,01431	0,99506
1,79	0,08038	0,963 27	2,19	0,036 26	0,98574	2,59	0,01394	0,99520
1,80	0,07895	0,96407	2,20	0,03547	0,986 10	2,60	0,01358	0,99534
1,81	0,077 54	0,96485	2,21	0,034 70	0,98645	2,61	0,01323	0,99547
1,82	0,07614	0,965 62	2,22	0,03394	0,98679	2,62	0,01289	0,99560
1,83	0,074 77	0,96638	2,23	0,03319	0,98713	2,63	0,01256	0,99573
1,84	0,073 41	0,96712	2,24	0,03246	0,98745	2,64	0,01223	0,99585
1,85	0,07206	0,96784	2,25	0,031 74	0,98778	2,65	0,01191	0,99598
1,86	0,07074	0,96856	2,26	0,031 03	0,98809	2,66	0,011 60	0,99609
1,87	0,069 43	0,96926	2,27	0,03034	0,98840	2,67	0,011 30	0,99621
1,88	0,06814	0,96995	2,28	0,02965	0,98870	2,68	0,01100	0,99632
1,89	0,066 87	0,970 62	2,29	0,02898	0,98899	2,69	0,01071	0,99643
1,90	0,06562	0,971 28	2,30	0,02833	0,98928	2,70	0,01042	0,99653
1,91	0,06439	0,97193	2,31	0,02768	0,98956	2,71	0,01014	0,99664
1,92	0,06316	0,97257	2,32	0,02705	0,98983	2,72	0,00987	0,99674
1,93	0,06195	0,97320	2,33	0,02643	0,99010	2,73	0,00961	0,99683
1,94	0,06077	0,97381	2 34	0 025 82	0 990 36	2,74	0,00935	0,9969
1,95	0,05959	0,97441	2,35	0,02522	0,990 61	2,75	0,009 09	0,997 01
1,96	0,05844	0,97500	2,36	0,02463	0,99086	2,76	0,00885	0,997 11
1,97	0,05730	0,97558	2,37	0,02406	0,991 11	2,77	0,00861	0,997 2(
1,98	0,056 18	0,97615	2,38	0,02349	0,991 34	2,78	0,00837	0,997 28
1,99	0,05508	0,976 70	2,39	0,02294	0,991 58	2,79	0,008 14	0,997 36
2,00	0,05399	0,977 2	2,40	0,02239	0,991 80	2,80	0,00792	0,9974
2,01	0,05292	0,977 7	2,41	0,021 86	0,992 0	2,81	0,007 70	0,997 5
2,02	0,051 86	0,978 31	2,42	0,021 34	0,9922	2,82	0,00748	0,9976
2,03	0,05082	0,9788	2,43	0,02083	0,9924	2,83	0,007 27	0,997 67
2,04	0,04980	0,979 3	2,44	0,02033	0,9926	2,84	0,00707	0,997 74

Продолжение табл. А1

x	$\phi(x)$	$\Phi(x)$	x	$\phi(x)$	$\Phi(x)$	x	$\phi(x)$	$\Phi(x)$
2,85	0,00687	0,997 81	3,25	0,00203	0,999 42	3,65	0,00051	0,99987
2,86	0,00668	0,99788	3,26	0,00196	0,99944	3,66	0,00049	0,999 87
2,87	0,00649	0,99795	3,27	0,001 90	0,99946	3,67	0,00047	0,999 88
2,88	0,00631	0,99801	3,28	0,001 84	0,99948	3,68	0,00046	0,999 88
2,89	0,00613	0,99807	3,29	0,001 78	0,99950	3,69	0,00044	0,999 89
2,90	0,005 95	0,99813	3,30	0,00172	0,99952	3,70	0,00042	0,99989
2,91	0,00578	0,99819	3,31	0,00167	0,99953	3,71	0,00041	0,99990
2,92	0,00562	0,99825	3,32	0,00161	0,99955	3,72	0,000 39	0,99990
2,93	0,00545	0,99831	3,33	0,001 56	0,99957	3,73	0,00038	0,99990
2,94	0,005 30	0,99836	3,34	0,001 51	0,99958	3,74	0,00037	0,99991
2,95	0,00514	0,99841	3,35	0,00146	0,99960	3,75	0,00035	0,99991
2,96	0,00499	0,99846	3,36	0,00141	0,99961	3,76	0,00034	0,99992
2,97	0,00485	0,99851	3,37	0,001 36	0,99962	3,77	0,00033	0,99992
2,98	0,004 71	0,99856	3,38	0,00132	0,99964	3,78	0,000 31	0,99992
2,99	0,00457	0,99861	3,39	0,001 27	0,99965	3,79	0,00030	0,999 92
3,00	0,004 43	0,99865	3,40	0,001 23	0,99966	3,80	0,000 29	0,999 93
3,01	0,00430	0,99869	3,41	0,001 19	0,99968	3,81	0,00028	0,999 93
3,02	0,004 17	0,998 74	3,42	0,001 15	0,99969	3,82	0,00027	0,99993
3,03	0,00405	0,99878	3,43	0,001 11	0,99970	3,83	0,00026	0,99994
3,04	0,00393	0,998 82	3,44	0,00107	0,99971	3,84	0,000 25	0,999 94
3,05	0,00381	0,99886	3,45	0,001 04	0,99972	3,85	0,00024	0,99994
3,06	0,00370	0,99889	3,46	0,00100	0,99973	3,86	0,00023	0,99994
3,07	0,00358	0,99893	3,47	0,00097	0,99974	3,87	0,00022	0,999 95
3,08	0,00348	0,99897	3,48	0,00094	0,99975	3,88	0,00021	0,99995
3,09	0,00337	0,99900	3,49	0,000 90	0,99976	3,89	0,00021	0,99995
3,10	0,00327	0,999 03	3,50	0,000 87	0,999 77	3,90	0,00020	0,99995
3,11	0,003 17	0,99906	3,51	0,00084	0,99978	3,91	0,00019	0,99995
3,12	0,00307	0,99910	3,52	0,00081	0,99978	3,92	0,00018	0,99996
3,13	0,002 98	0,99913	3,53	0,00079	0,999 79	3,93	0,00018	0,99996
3,14	0,00288	0,99916	3,54	0,00076	0,99980	3,94	0,00017	0,99996
3,15	0,00279	0,999 18	3,55	0,00073	0,99981	3,95	0,00016	0,99996
3,16	0,00271	0,99921	3,56	0,000 71	0,99981	3,96	0,00016	0,99996
3,17	0,00262	0,99924	3,57	0,00068	0,99982	3,97	0,000 15	0,999 96
3,18	0,00254	0,999 26	3,58	0,00066	0,999 83	3,98	0,00014	0,99997
3,19	0,00246	0,999 29	3,59	0,000 63	0,99983	3,99	0,00014	0,99997
3,20	0,00238	0,99931	3,60	0,00061	0,99984			
3,21	0,00231	0,99934	3,61	0,00059	0,99985			
3,22	0,00224	0,999 36	3,62	0,00057	0,99985			
3,23	0,00216	0,99938	3,63	0,00055	0,999 86			
3,24	0,002 10	0,99940	3,64	0,000 53	0,99986			

Таблица А.2. Верхние ЮОа %-ные точки * распределения χ^2. Входом таблицы служит x, для которого $P\left(\chi^2_\nu > x\right) = \alpha$



	α.									
ν	0,995	0,99	0,975	0,95	0,9	0,1	0,05	0,025	0,01	0,005
1	$0{,}0^4393$	$0{,}0^3157$	$0{,}0^3982$	$0{,}0^2393$	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,61	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34

* Нижняя ЮОа %-ная точка равна верхней $100(1-\alpha)$ %-ной точке.

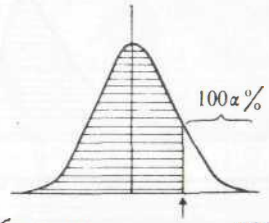
Источник. Pearson E. S., Thompson C. M. Tables of the Percentage Points of the Incomplete Beta Function and of the Chi-square Distribution. Biometrika, vol. 32 (1941). Воспроизведено с разрешения издателя.

Продолжение табл. А.2

	a									
v	0,995	0,99	0,975	0,95	0,9	0,1	0,05	0,025	0,01	0,005
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,4	104,2
80	51,1.7	53,54	57,15	60,39	64,28	96,58	101,9	106,6	112,3	116,3
90	59,20	61,75	65,65	69,13	73,29	107,6	113,1	118,1	124,1	128,3
100	67,33	70,06	74,22	77,93	82,36	118,5	124,3	129,6	135,8	140,2

Замечание. Для больших значений v необходимо воспользоваться аппроксимацией (см. с. 99).

Таблица А.3. Верхние 100α %-ные точки* t-распределения Стьюдента. Входом таблицы служит x, для которого $P(t_\nu > x) = \alpha$

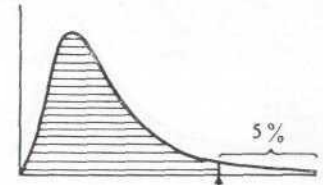


	a					
v	0,005	0,01	0,025	0,05	0,1	0,15
1	63,657	31,821	12,706	6,314	3,078	1,963
2	9,925	6,965	4,303	2,920	1,886	1,386
3	5,841	4,541	3,182	2,353	1,638	1,250
4	4,604	3,747	2,776	2,132	1,533	1,190
5	4,032	3,365	2,571	2,015	1,476	1,156
6	3,707	3,143	2,447	1,943	1,440	1,134
7	3,499	2,998	2,365	1,895	1,415	1,119
8	3,355	2,896	2,306	1,860	1,397	1,108
9	3,250	2,821	2,262	1,833	1,383	1,100
10	3,169	2,764	2,228	1,812	1,272	1,093
11	3,106	2,718	2,201	1,796	1,363	1,088
12	3,055	2,681	2,179	1,782	1,356	1,083
13	3,012	2,650	2,160	1,771	1,350	1,079
14	2,977	2,624	2,145	1,761	1,345	1,076
15	2,947	2,602	2,131	1,753	1,341	1,074
16	2,921	2,583	2,120	1,746	1,337	1,071
17	2,898	2,567	2,110	1,740	1,333	1,069
18	2,878"	2,552	2,101	1,734	1,330	1,067
19	2,861	2,539	2,093	1,729	1,328	1,066
20	2,845	2,528	2,086	1,725	1,325	1,064
21	2,831	2,518	2,080	1,721	1,323	1,063
22	2,819	2,508	2,074	1,717	1,321	1,061
23	2,807	2,500	2,069	1,714	1,319	1,060
24	2,797	2,492	2,064	1,711	1,318	1,059
25	2,787	2,485	2,060	1,708	1,316	1,058
26	2,779	2,479	2,056	1,706	1,315	1,058
27	2,771	2,473	2,052	1,703	1,314	1,057
28	2,763	2,467	2,048	1,701	1,313	1,056
29	2,756	2,462	2,045	1,699	1,311	1,055
30	2,750	2,457	2,042	1,697	1,310	1,055
∞	2,576	2,326	1,960	1,645	1,282	1,036

* Для получения нижних 100α %-ных точек необходимо поменять знак у верхней 100α %-ной точки.

Источник. Fisher R. A. Statistical Methods for Research Workers. Messrs. Oliver and Boyd. Воспроизводится с разрешения автора и издателя.

Таблица А.4а. **Верхние 5 %-ные точки* F-распределения. Входом таблицы служит x, для которого $P(F_{m,n}>x)=0{,}05$**



Табличное значение критерия

n \ m	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,19
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

n \ m	12	15	20	24	30	40	60	120	его	n
1	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3	1
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50	2
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53	3
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63	4
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36	5
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67	6
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23	7
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93	8
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71	9
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54	10
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40	11
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30	12
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21	13
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13	14
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07	15
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01	16
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96	17
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92	18
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88	19
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84	20
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81	21
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78	22
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76	23
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73	24
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71	25
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69	26
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67	27
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65	28
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64	29
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62	30
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51	40
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39	60
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25	120
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00	∞

* Нижняя 5 %-ная точка $F_{m,n}$-распределения равна величине, обратной к его верхней 5 %-ной точке.

Источник. [76, табл. 18]. Воспроизведено с разрешения Biometrika Trustees.

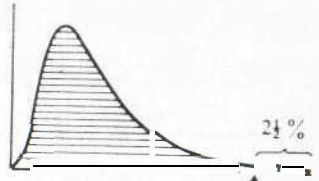
Таблица А.4б. Верхние 2,5 %-ные точки* F-распределения. Входом таблицы

n \ m	1	2	3	4	5	6	7	8	9	10
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
16	6,12	4,69.	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99
17	6,04	4,62	4.01	3,66	3,44	3,28	3,16	3,06.	2,98	2,92
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67
24	5,72	4,32	3,7?	3,38	3,15	2,99	2,87	2,78	2,70	2,64
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16
∞	5,02	3,69	3,12	2,79	2,57	2,4!	2,29	2,19	2,11	2,05

* Нижняя 2,5 %-ная точка $F_{m,n}$-распределения равна величине, обратной к его верхней 2,5 %-ной точке.

Источник. [76, табл. 18]. Воспроизведено с разрешения Biometrika Trustees.

служит x, для которого $P(F_{m,n} > x) = 0{,}025$



n \ m	12	15	20	24	30	40	60	120	∞
1	976,7	984,9	993,1	997,2	1001	1006	1010	1014	1018
2	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
3	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
4	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

2,5 %-ной точке.

ЛИТЕРАТУРА

1. Anscombe F. J. The transformation of Poisson, binomial and negative binomial data. Biometrika, vol. 35, p. 246—254, 1948.
2. Bailey N. T. J. Statistical Methods in Biology. English Universities Press, 1959. Русский перевод: Бейли Н. Статистические методы. М., Мир, 1963.
3. Balaam L. N. Fundamentals of Biometry. George Allen and Unwin Ltd, 1972.
4. Bortkiewicz L. von. Das Gesetz der Kleinen Zahlen. Leipzig: Teubner, 1898.
5. Box G. E. P. Non-normality and tests on variances. Biometrika, vol. 40, p. 318—335, 1953.
6. Box G. E. P. and Müller M. E. A note on the generation of random normal deviates. Annals of Mathematical Statistics, vol. 29, p. 610—611, 1958.
7. Brownlee K. A. Statistical Theory and Methodology in Science and Engineering (second edition), Wiley, 1965. Русский перевод: Браунли К. Статистическая теория и методология в науке и технике. М., Наука, 1977.
8. Campbell R. C. Statistics for Biologists. Cambridge University Press, 1974.
9. Chakravarti I. M., Laha R. G. and Roy J. Handbook of Methods of Applied Statistics, vol. 1, Wiley, 1967.
10. Chakravarti I. M., Laha R. G. and Roy J. Handbook of Methods of Applied Statistics, vol. 2, Wiley, 1967.
11. Chemical Rubber Publishing Company Standard Mathematical Tables (C. D. Hodgman, Editor), 1963.
12. Clarke R. D. An Application of the Poisson distribution. J. Institute of Actuaries, vol. 72, p. 481, 1946.
13. Cochran W. G. Some methods for strengthening the common χ^2 tests. Biometrics, vol. 10, p. 417—451, 1954.
14. Cohen A. C. Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. Annals of Mathematical Statistics, vol. 21, p. 557—569, 1950.
15. Cohen A. C. On the solution of estimating equations for truncated and censored samples from normal populations. Biometrika, vol. 44, p. 225—236, 1957.
16. Colquhoun D. Lectures on Biostatistics. Clarendon Press, 1971.
17. Conte S. D. Elementary Numerical Analysis. McGraw-Hill, 1965.
18. Craig C. C. On the utilisation of marked specimens in estimating populations of flying insects. Biometrika, vol. 40, p. 170—176, 1953.
19. Draper N. R. and Smith H. Applied Regression Analysis. Wiley, 1956. Русский перевод: Дрейпер Н. и Смит Г. Прикладной регрессионный анализ. М., Статистика, 1973.
20. Elderton W. P. and Johnson N. L. Systems of Frequency Curves. Cambridge University Press, 1969.
21. Feller W. An Introduction to Probability Theory and Its Applications, vol. 1, Wiley, 1950. Русский перевод: Феллер В. Введение в теорию вероятностей и ее приложения, т. 1. М., Мир, 1967.
22. Ferguson R. A., Fryer J. G. and McWhinney I. A. On the estimation of a truncated normal distribution. Contributed paper to the International Statistical Institute meeting in Warsaw, 1975.
23. Fisher R. A. Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. J. Agricultural Science, vol. 11, p. 107—135, 1921.
24. Fisher R. A., Corbet A. S. and Williams C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Animal Ecology, vol. 12, p. 42—57, 1943.
25. Fisher R. A. and Yates F. Statistical Tables for Use in Biological, Agricultural and Medical Research (sixth edition). Oliver and Boyd, 1963.
26. Fraser D. A. S. Statistics — An Introduction. Wiley, 1967.
27. Freeman H. Finite Differences for Actuarial Students. Cambridge University Press, 1967.
28. Fröberg C. Introduction to Numerical Analysis. Addison— Wesley, 1965.
29. Gauss K. F. Theoria motus corporum coelestium in sectionibus conicus solem ambientium auctore C. F. G. 1809.
30. Gauss K. F. Theoria combinationis observationum erroribus minimis obnoxiae. Gottingen, 1823.
31. Gauss K. F. Supplementum theoriae combinationis observationum erroribus minimis obnoxiae. Gottingen, 1828.
32. Gauss K. F Methode des Moindres Carres (French translation by J. Bertrand). Paris, 1855.
33. Gauss K. F. Theory of the Motion of the Heavenly Bodies, moving about the Sun in Conic Sections: a Translation of Gauss' "Theoria Motus" with an Appendix by Charles Henry Davis. Boston, 1857.

34. Glasser G. J. and Winter R. F. Critical values of rank correlation for testing the hypothesis of independence. Biometrika, vol. 48, p. 444—448, 1951.
35. Greville T. N. E. (1947). Actuarial note: adjusted average graduation formulas of maximum smoothness. The Record, American Institute of Actuaries, vol. 36, p. 249—261, 1947.
36. Greville T. N. E. Actuarial note: tables of coefficients in adjusted average graduation formulas of maximum smoothness. The Record, American Institute of Actuaries, vol. 37, p. 11—30, 1948.
37. Greville T. N. E. Theory and Application of Spline Functions. Academic Press, 1969.
38. Grossman S. I. and Turner J. E. Mathematics for the Biological Sciences. Macmillan, 1974.
39. Guenther W. C. Analysis of Variance. Prentice-Hall, 1964.
40. Hammersley J. M. and Handscomb D. C. Monte Carlo Methods. Methuen, 1967.
41. Hartree D. R. Numerical Analysis, Oxford University Press, 1955.
42. Henrici P. Elements of Numerical Analysis. Wiley, 1964.
43. Hildebrand F. B. Introduction to Numerical Analysis. McGraw-Hill, 1956.
44. Hocking R. R. The analysis and selection of variables in linear regression. Biometrics, vol. 32, p. 1—50, 1976.
45. Hoel P. G. Elementary Statistics. Wiley, 1971.
46. Hoel P. G. Introduction to Mathematical Statistics. Wiley, 1971.
47. Johnson N. L. and Kotz S. Distributions in Statistics (vol. 1). Discrete Distributions. Houghton Mifflin. Boston, 1969.
48. Johnson N. L. and Kotz S. Distributions in Statistics (vol. 2). Continuous Univariate Distributions 1. Houghton Mifflin. Boston, 1970.
49. Johnson N. L. and Kotz S. Distributions in Statistics (vol. 3). Continuous Univariate Distributions 2. Houghton Mifflin, Boston, 1970.
50. Johnson N. L. and Leone F. C. Statistics and Experimental Design in Engineering and the Physical Sciences, vol. 1, Wiley, 1964. Русский перевод: Джонсон Н. и Лион Ф. Статистика и планирование эксперимента в технике и науке. М., Мир, 1980 (т. 1).
51. Johnson N. L. and Leone F. C. Statistics and Experimental Design in Engineering and the Physical Sciences, vol. 2, Wiley, 1964.
52. Kendall M. G. Rank correlation Methods. Griffin, 1955. Русский перевод: Кендалл М. Ранговые корреляции. М., Статистика, 1975.
53. Kendall M. G. and Stuart A. The Advanced Theory of Statistics (vol. 1). Distribution Theory (second edition). Griffin, 1963. Русский перевод: Кендалл М. и Стьюарт А. Теория распределений. М., Наука, 1966.
54. Kendall M. G. and Stuart A. The Advanced Theory of Statistics (vol. 2). Inference and Relationship (second edition). Griffin, 1967. Русский перевод: Кендалл М., Стьюарт А. Статистические выводы и связи. М., Наука, 1973.
55. Kendall M. G. and Stuart A. The Advanced Theory of Statistics (vol. 3). Design and Analysis and Time Series (second edition). Griffin, 1968. Русский перевод: Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. М., Наука, 1976.
56. Kruskal W. H., Wallis W. A. Use of ranks in one-criterion variance analysis. J. American Statistical Association, vol. 47, p. 583—621, 1952.
57. Lancaster H. O. The Chi-squared Distribution, Wiley, 1969.
58. Legendre A. M. Nouvelles méthodes pour la determination des orbites des comètes; avec un supplement. Paris. 1806.
59. Liebermann G. J., Owen D. B. Tables of the Hypergeometric Probability Distribution. Stanford University Press, 1961.
60. Lotka A. J. Population analysis — the extinction of families — I. J. Washington Academy of Sciences, vol. 21, p. 377—380, 1931.
61. Lotka A. J. Population analysis — the extinction of families — II. J. Washington Academy of Sciences, vol. 21, p. 453—459, 1931.
62. Lyon A. J. Dealing with data. Pergamon, 1970.
63. McCornack R. L. Extended tables of the Wilcoxon matched pair signed rank statistic. J. American Statistical Association, vol. 60, p. 864—871.
64. Massey F. J. The Kolmogorov — Smirnov test of goodness of fit. J. American Statistical Association, vol. 46, p. 68—78, 1951.
65. Mather K. Statistical Analysis in Biology. Chapman and Hall, 1965.
66. Mather K. The Elements of Biometry. Chapman and Hall, 1971.
67. Miller M. D. Elements of Graduation. Actuarial Society of America. 1946.
68. Milton R. C. Extended table of critical values for the Mann—Whitney (Wilcoxon) two-sample statistics. J. American Statistical Association, vol. 59, p. 925—934, 1964.
69. Molina E. C. Poisson's Exponential Limit. Van Nostrand, 1942.
70. Mood A. M., Graybill F. A. Introduction to the Theory of Statistics. Mc-Graw-Hill, 1963.
71. National Bureau of Standards. Tables of the Binomial Probability Distribution. US Government Printing Office, 1950.
72. Neave H. R. On using the Box-Müller transformation with multiplicative congruential pseudo-random number generators. Applied Statistics, vol. 22, p. 92—97, 1973.
73. Nielsen K. L. Methods in Numerical Analysis, Macmillan, 1956.
74. Olds E. G. Distribution of sums of squares of rank differences for small numbers of individuals. Annals of Mathematical Statistics, vol. 9, p. 133—148, 1938.
75. Parzen E. Stochastic Processes, Holden-Day, 1962.
76. Pearson E. S., Hartley H. O. Biometrika Tables for Statisticians, vol. 1. Cambridge University Press, 1958.
77. Pearson E. S., Hartley H. O. Biometrika Tables for Statisticians, vol. 2. Cambridge University Press, 1972.
78. Peters J. A. (ed.) Classic Papers in Genetics. Prentice-Hall, 1959.

79. Pielou E. C. An Introduction to Mathematical Ecology, Wiley-Interscience, 1969.
80. Plackett R. L. A historical note on the methods of least squares. Biometrika, vol. 36, p. 458—460, 1949.
81. Pollard A. H. Introductory Statistics — A Service Course. Pergamon. 1972.
K. Pollard J. H. On distance estimators of density in randomly-distributed forests. Biometrics, vol. 27, p. 991—1002, 1971.
S3. Pollard J. H. Mathematical Models for the Growth of Human Populations. Cambridge University Press, 1973.
84. Ralston A. A first Course in Numerical Analysis. McGraw-Hill, 1965.
85. Rand Corporation. A Million Random Digits and 100.000 Normal Deviates. The Free Press. Glencoe. Illinois, 1955.
86. Remington R. D., Schork M. A. Statistics with Applications to the Biological and Health Sciences. Prentice-Hall, 1970.
87. Samiuddin M., Atiqullah M. A test of equality of variances. Biometrika, vol. 63, p. 206—208.
88. Scheffé H. Analysis of Variance. Wiley, 1959. Русский перевод: Шеффе Г. Дисперсионный анализ. М., Наука, 1980.
89. Scheid F. Numerical Analysis. Schaum's Outline Series. McGraw-Hill, 1938.
90. Skellam J. G. Studies in statistical ecology. Biometrika, vol. 39, p. 346—362, 1952.
91. Snedecor G. W. Statistical Methods (fifth edition), Iowa State University Press, 1961. Русский перевод: Снедекор Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. М., Сельхозиздат, 1961.
92. Sokolnikoff I. S., Sokolnikoff E. S. Higher Mathematics for Engineers and Physicists. McGraw-Hill, 1941.
93. Spiegel M. R Theory and Problems of Statistics. Schaum's Outline Series McGraw-Hill. 1972.
94. 'Student'. On the probable error of a mean. Biometrika. vol. 6. pp. 1—25, 1908a.
95. 'Student'. On the probable error of a correlation coefficient. Biometrika, vol. 6. p. 302—310, 1908b.
96. Tetley H. Actuarial Statistics, vol. 1. Statistics and Graduation, Cambridge University Press, 1966.
97. Thöni H. Transformation of Variables Used in the Analysis of Experimental and Observational Data. A Review. Technical Report Number 7. Statistical Laboratory, Iowa State University. Ames, Iowa, 1967.
98. U. S. Army Material Command. Engineering Design Handbook. Tables of the Cumulative Binomial Probabilities, 1972.
99. Weintraub S. Tables of the Cumulative Binomial Probability Distribution for Small Values of p. The Free Press of Glencoe, 1963.
100. Whittaker E., Robinson G. The Calculus of Observations. Blackie. Русский перевод: Уиттаккер Э. Т., Робинсон Г. Математическая обработка результатов наблюдений. М., ОНТИ, 1935.
101. Wilks S. S. Mathematical Statistics. Wiley, 1962. Русский перевод Уилкс С. С. Математическая статистика. М., Наука, 1967.
102. Wilks S. S. Elementary Statistical Analysis. Princeton University Press, 1966.
1C3. Williams C. B. Patterns in the Balance of Nature. Academic Press. 1964.
104. Williamson E., Bretherton M. H. Tables of the Negative Binomial Probability Distribution. Wiley, 1963.
105. Zar J. H. Biostatistical Analysis. Prentice-Hall, 1974.

СОДЕРЖАНИЕ

Дж. Поллард

СПРАВОЧНИК ПО ВЫЧИСЛИТЕЛЬНЫМ МЕТОДАМ СТАТИСТИКИ

Книга одобрена на заседании секции редсовета издательства 26.09.79

Зав. редакцией *А. В. Павлюков*
Редактор *Е. В. Крестьянинова*
Мл. редактор *О. Б. Степанченко*
Техн. редактор *И. В. Завгородняя*
Корректоры *Г. В. Хлопцева, Т. М. Васильева* и *А. Т. Сидорова*
Худож. редактор *О. Я. Поленова*
Переплет художника *Б. С. Вехтера*

ИБ № 1017

Сдано в набор 04.02.82. Подписано в печать 03.08.82. Формат 60×90¹/₁₆. Бум. тип. № 1. Гарнитура «Литературная». Печать высокая. П. л. 21,5. Усл. п. л. 21,5. Уч.-изд. л. 22,28 Тираж 15000 экз. Заказ 160. Цена 1 р. 60 к.
Издательство «Финансы и статистика», Москва, ул. Чернышевского, 7.

Ленинградская типография № 8 ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 190000, г. Ленинград. Прачечный переулок, 6.